POLITECNICO DI MILANO

School of Civil, Environmental and Land Planning Engineering

ÉCOLE CENTRALE DE LYON

CENTRE LÉON BÉRARD

Comparative assessment of different modeling approaches in estimating atmospheric exposures to air pollution for epidemiological studies on cancer

Master of Science in Environmental and Land Planning Engineering

Academic year 2019/2020

Author

Riccardo MOTTALINI

Supervisors

Giovanni LONATI

Pietro SALIZZONI

Thomas COUDON

Acknowledgements

I would like to thank Professor Giovanni Lonati and Professor Pietro Salizzoni for the opportunity they gave me to carry out such an inspiring experience, for their support and their valuable teachings about atmospheric pollution.

I express my sincere gratitude to my patient and supportive *maître de stage* Thomas Coudon, for his constant guidance and encouragement during all the running of this project and for his remarkable efforts with a (sometimes) lackluster French speaker.

A special thanks is also due to Lény Grassot, Matheiu Dubuis and Delphine Praud, whose great knowledge of environmental epidemiology and continuous assistance have been extremely valuable in the development of this work. I am also very grateful to the colleagues and researchers with whom I spent these months in Lyon, both for the academic advice and for their help in getting to know the city and settling in. A special thanks to Alessandro De Giovanni, roommate and fellow traveler.

Finally, I cannot forget to thank my family, my girlfriend Marta and all my friends for the love and support they gave me in these rewarding years of university.

Riccardo Mottalini

Abstract

The objective of this study was the relative comparison of different modeling approaches in estimating the air pollution exposure for epidemiological studies on cancer. Values estimated by a dispersion model (SIRANE), a LUR model and other less elaborated approaches (regional model CHIMERE, interpolation and proximity models) on 785 members of the E3N cohort within the Lyon Metropolitan Area were compared, as well for multiple virtual populations. The pollutant considered were NO_2, O_3 and PM_{10} and the comparison was set between annual average values for the year 2010 and 2000. The land use type and a socioeconomic factor (average income) were investigated as possible sources of misclassifications between SIRANE and LUR. The variation in odds ratio (theoretical risk of breast cancer due to the exposure to NO_2 and PM_{10}) was assessed by replacing SIRANE exposure value by LUR exposure value over 10 000 virtual subjects, with an iterative procedure.

Good correlation was observed between SIRANE and LUR (r >0.7, ρ >0.8, $w\kappa$ >0.6) for all pollutants both in 2010 and 2000, while other models demonstrated lower agreement. The LUR model showed a tendency to overestimate PM_{10} and not to capture fine-scale ozone spatial variability. Furthermore, it was observed that LUR overestimate NO_2 within a continuous urban fabric. The epidemiological outputs comparison indicated that LUR slightly underestimate odds ratios with respect to SIRANE, potentially leading to mis information in breast cancer risk estimation. Considering highly exposed populations, the loss of significance was important.

The LUR model has been evaluated as a good alternative to a dispersion model in estimating exposure values for epidemiological studies, although showing an inferior capacity to capture small-scale variation that make it less feasible to studies focusing only on populations exposed to a small range of concentration values.

Abstract ITA

Il presente studio ha avuto come obiettivo il confronto di differenti approcci modellistici per la stima dell'esposizione all'inquinamento urbano in studi epidemiologici sul cancro. Sono stati confrontati i valori di esposizione stimati da un modello di dispersione (SIRANE), un modello land use regression (LUR) e altri approcci (modello CHIMERE, modello "Nearest-AQMS", modelli di prossimità) su 785 membri della coorte "E3N", residenti nella zona metropolitana della città di Lione, e su altre popolazioni create virtualmente. Gli inquinanti considerati sono stati NO_2, PM_{10} e O_3 e il confronto è stato impostato tra i valori medi annuali per gli anni 2010 e 2000. Per quanto riguarda SIRANE e il modello LUR, il tipo di uso del suolo e fattori socioeconomici (reddito medio) sono stati valutati come possibili fonti di disaccordo tra i due modelli mediante analisi geografiche. Le differenze nel calcolo degli odds ratio (rischio teorico di cancro al seno legato all'esposizione all'inquinamento) tra i due modelli è stata valutata sostituendo ai valori SIRANE con quelli stimati da LUR su 10000 soggetti, applicando una procedura iterativa.

Lo studio ha mostrati buoni livelli di correlazione per SIRANE-LUR (r >0.7, ρ >0.8, $w\kappa$ >0.6) sia per il 2010 che per il 2000, mentre gli altri modelli hanno dato risultati peggiori. Il modello LUR ha mostrato una tendenza a sovrastimare il PM_{10} e a non descrivere precisamente la variabilità spaziale dell'ozono. Inoltre, si è osservato che il modello LUR sovrastima l'NO_2 nel tessuto urbano continuo. La stima degli output epidemiologici ha indicato che il modello LUR sottostima leggermente gli odds ratio rispetto a SIRANE, causando una sottostima del rischio di cancro . Considerando popolazioni fortemente esposte, si è osservata una perdita di significatività epidemiologica importante.

Il modello LUR è stato valutato come una valida alternativa ai modelli di dispersione per quanto riguarda la stima dell'esposizione per fini epidemiologici, mostrando comunque un'inferiore capacità di descrivere variazioni a piccola scala che lo rende meno adatto a studi con popolazioni esposte solamente a range di concentrazione ristretti.

Contents

List of Figures

List of Tables

1 Introduction

1.1 The burden of air pollution and its principal effects on human health

Outdoor air pollution is a major public health problem leading to adverse health effect The World Health Organization defined in 2016 air pollution as the biggest environmental risk in to health, underlying that 90% of people in the world breath air that does not comply with the WHO Air Quality Guidelines [WHO, 2016]. In 2015, air pollution related diseases were responsible for about 6.4 million premature deaths, with 4.2 million due to ambient air pollution and 2.8 million to indoor pollution, being more than 10% of all worldwide deaths.

Lelieveld et al published in 2020 a study that compares the loss of life expectancy from air pollution with other risk factors, such as tobacco smoking and AIDS [Lelieveld et al., 2020]: for air pollution, a global LLE (Loss of Life Expectancy) of 2.9 years was observed, higher than those found for tobacco smoking and AIDS, respectively 2.2 and 0.7.

The YLL (Years of Life Lost) spatial distribution all over the world (figure 1.1) shows that air pollution impact almost every country in the world, with Central Asia, East Asia as the most affected regions.

From the perspective of environmental justice, the reduction of air pollution is also key, since it is recognized that the poorest and most vulnerable people are the most affected by pollution (92% of pollution-related deaths occur in low-income and middle-income countries [Landrigan et al., 2018]). The number of deaths is projected to increase without

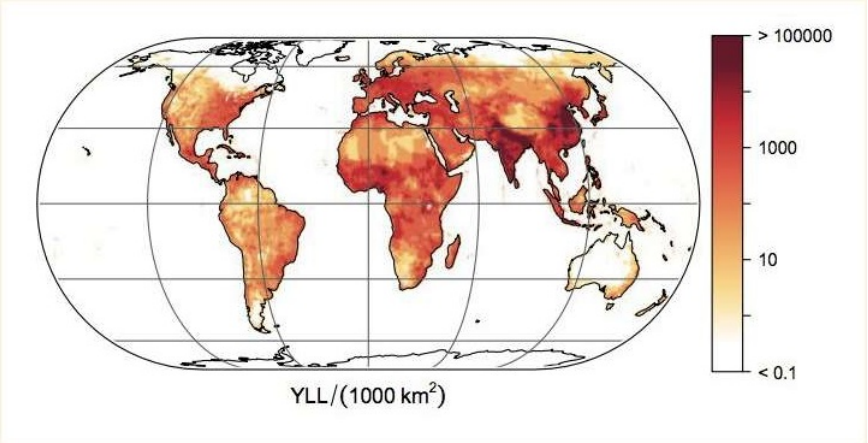

Figure 1.1: Annual years life lost from air pollution all over the world [Lelieveld et al., 2020]

strong politic strategies, mainly as a result of the exponential growth of cities and of energy demand in developing countries.

Furthermore, high-risks of pollution-related diseases are also related to children during periods of great vulnerability in pregnancy and in early infancy, most of the times in terms of respiratory infections and childhood asthma [WHO, 2017],[Landrigan et al., 2019]. The most frequents causes of morbidity and mortality due to long-term exposure to air pollution are those related to NCDs (Non-Communicable-Diseases, chronic diseases of long duration [Forouzanfar et al., 2016]), principally of respiratory and cardiovascular nature. Most frequent NCDs due to air pollution are chronic obstructive pulmonary disease, ischemic heart disease, lung cancer and lower respiratory infections (figure 1.2).

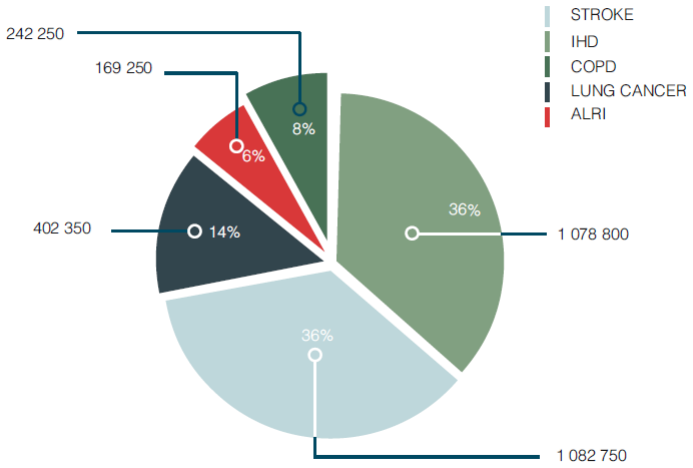

Figure 1.2: Percentages of total ambient air pollution burden in 2012; ALRI: acute lower respiratory disease; COPD: chronic obstructive pulmonary disease; IHD: ischemic heart disease [WHO, 2016]

1.2 Environmental epidemiology

Air pollution is composed by a huge variety of chemical species and derives from several sources. Adverse effect on health given by different pollutant are consequently extremely complex and not properly understood [Bernstein et al., 2004]. The risk assessment for air pollution exposure is usually conducted investigating statistical associations between air pollutant levels and various outcomes, such as the number of hospital admissions. Epidemiology is the science that studies distributions, patterns and causes of diseases in a defined population, identifying risk factors for public health and providing targets for health care strategies. The branch of epidemiology that is in charge of determining how environmental exposures impact on human health is called environmental epidemiology.

To improve public health and develop prevention policies, environmental exposures of the population are increasingly being monitored through measurements and simulations. In this context, studies on the links between air pollution and health risks are increasing. On the other hand, this movement is supported by a better reduction of environmental risks for the general public following the appearance of several environmental disasters in Europe, the US and the world since the 1960s.

Most of these analyses are carried out in urban environment, which is usually associated with high concentrations of outdoor air pollutants due to the great number of pollutant sources that are distinctively proper of densely populated areas [Ezzati and Organització Mundial de la Salut, 2004]. In urban environments and especially in those areas where population and traffic density are relatively high (near busy traffic axis in city center), the urban topography and the urban microclimate contribute to develop poor air dispersion conditions and create concentration hotspots [Vardoulakis et al., 2003]. Since the world population is becoming more and more urbanized (fully half of the world's population now live in urban areas [Gilbert and Wendell, 2014]), epidemiological studies in urban environment are increasingly developing to support public authorities in the decision making process.

Urban air pollution is mainly due to combustion processes and its major sources are traffic (mobile sources), industrial processes and building heating (stationary sources), emitting into the atmosphere a complex mixture of pollutants, that could vary depending on the relative contribution of different sources and on the effect of climatic factors. Most frequently and routinely monitored air pollutant include particulate matter (PM), nitrogen dioxide (NO_2) and ozone (O_3). Others are carbon oxide, lead, black smoke and soot.

PM_{10}

Atmospheric particulate matter consists of any dispersed matter, solid or liquid, in which the individual aggregates range from molecular clusters of 0.005 μm diameter to coarse particles up to about 100 μm. Particulate matter (PM) can be emitted directly as carbonaceous soot particles from incomplete combustion, or it can be formed into the atmosphere (for example when gaseous NO_x and SO_2 are transformed through heterogeneous reactions in sulfates or nitrates). Although particles may have a very irregular shape, their size can be described by an equivalent *aerodynamic diameter* determined by comparing them with perfect spheres having the same setting velocity. The particles of most interest have aerodynamic diameter in the range of 0.1 μm up to 10 μm. Particles smaller than 2.5 μm ($PM_{2.5}$) are referred to as *fine* particles, while PM_{10} refers to all particulate matter with aerodynamic diameter below 10 μm. The impact of PM_{10} on human health is strictly correlated at the size of the particles inhaled. Larger particles that enters the respiratory system can be trapped by the hairs and lining of the nose and then cough off. Smaller particles that arrive to the tracheobronchial system can be captured by mucus or other defense mechanisms but may also be able to traverse it and deposit into the lungs. Particulates aggravate existing respiratory and cardiovascular diseases and damage lung tissue. Additionally, due to their nature, some are carcinogenic. Associations between exposure to PM and cancer occurrence has been observed my multiple studies [Andersen et al., 2017a],

[Raaschou-Nielsen et al., 2016], [Weinmayr et al., 2018].

NO_2

Among the several oxides of nitrogen that are known to occur, only NO and NO_2 are important air pollutants. There are two sources of nitrogen oxides (also named NO_x) when fossil fuels are burned: thermal NO_x, which are created when nitrogen and oxygen in combustion air are heated to very high temperature (>1000 K), and fuel NO_x, which results from the oxidation of nitrogen compounds chemically bound in the fuel molecules. 95% of anthropogenic emissions of NO_x are in form of NO, a colorless gas that has no known adverse effects on human health. However, NO readily oxidize to NO_2, that can irritate lungs, cause bronchitis and pneumonia and lower the system resistance to respiratory infections [Kampa and Castanas, 2008]. NO_2 has also been linked to breast cancer development in a meta-analysis of individual data from 15 European cohorts [Andersen et al., 2017b]. Other consequences due to NO_x presence in air are its reactions with volatile organic compounds in presence of sunlight to form photochemical oxidants that have adverse health effects as well.

O_3

The simultaneous presence of organic compounds, NO_x and sunlight can initiate a complex set of reactions that produce a number of secondary pollutants known as photochemical oxidants, of which Ozone (O_3) is the most abundant. Ozone pollution is therefore mainly associated with warmer months, when the weather conditions that favor the formation of ground-level ozone are present. O_3 in ambient air has been associated with a variety of transient effects on the human body, namely asthma, bronchitis, heart attack and other cardiopulmonary problems. Furthermore, long-term exposure to ozone has been shown to increase risk of death from respiratory illness: a study of 450000 people living in United States cities saw a significant correlation between ozone levels and respiratory illness over a 18-year follow-up period, revealing that people living in cities with

high ozone levels had an over 30% increased risk of dying from lung disease [Jerrett et al., 2009]. One of the main characteristics of ozone is that higher surface O_3 concentrations are measured in rural areas than in urban areas because ozone levels are higher downwind of its precursors' sources at distances of hundreds of kilometers [Monks et al., 2015].

1.3 Risk of breast cancer associated with ambient air pollution exposure: the XENAIR project

Breast cancer and air pollution

Global cancer statistics estimated that in 2018 breast cancer (BC) was the most common among women, with 2.09 million new cases diagnosed in the world [Bray et al., 2018]. In France, its incidence has continuously increased: this has been associated with mass screening, menopausal hormonal therapy but also with societal changes impacting lifestyles. A crucial role of lifestyle and environmental factors on the occurrence of BC has been suggested by epidemiological studies [Harvie et al., 2015] [Jemal et al., 2010], including ambient air pollution. However, the fact that exposure to environmental pollutants may play a role in BC development has been supported by both epidemiological and scientific findings [Brody et al., 2007] and is now evidenced. In 2013 the International Agency for Research on Cancer (IARC) classified the outdoor pollution as a whole (as well as PM) as carcinogenic to humans, principally based on studies on lung and bladder cancers [Loomis et al., 2014].

Epidemiological findings suggested associations between breast cancer occurrence and NO_2 from traffic-related air pollution [Nie et al., 2007]. Moreover, it has been reported that women with extremely dense mammography density, that is a proved risk factor for breast cancer, were less likely to have high levels of exposure to ozone [Yaghjyan et al., 2017]. Other pollutant species that were linked with BC are PCBs, benzo[α]pyrene, cadmium, PAHs [Amadou et al., 2019].

Main limitations of epidemiological studies are the lack of information about confounding personal risk factors (smoking, body weight, familiar cancer history, eating habits) and about pollutant to which subjects are exposed (often, only one pollutant or source is considered). Furthermore, the retrospective exposure reconstruction is made difficult by the lack of his-

torical exposure measures. Studies suggested that this limitation could be overcame by considering the urban residence as a surrogate for air pollution exposure due to urban sources, in order to investigate periods where historical air pollution records are unavailable [Binachon et al., 2014]. Another limit of most of studies on BC development is the consideration of adulthood exposures within short observation periods, while the exposure occurring during biological time windows of greater sensibility (during childhood, in utero) have been suggested to be more strongly correlated with BC risk [Potischman and Troisi, 1999]

XENAIR project

The XENAIR project is an interdisciplinary research project involving 6 different *équipes*, focusing on epidemiology, expology, geography and biostatistics:

- Département Cancer et Environnement, Centre Léon Bérard, Lyon
- Équipe Générations et santé - Inserm UMR 1018
- Équipe AIR, LMFA, École Centrale de Lyon, Écully, France
- INERIS, Verneuil-en-Halatte, Oise, France
- Leicester University, Center for Environment, Sustainability and Health, UK
- ISPED, Université de Bordeaux, France

The objective of the XENAIR project is to investigate chronic long-term effects of the exposure to multiple ambient air pollutants and risk of breast cancer in a nested case-control study within the E3N (Étude Épidémiologique auprès de femmes de la Mutuelle Générale de l'Éducation Nationale) cohort. Selected ambient air pollutants are PM, NO_2, O_3, benzo[α]pyrene, dioxins, PCB and Cd.

The study analyzes trajectory profiles of individual exposure values over time since recruitment, estimating BC risk associated with their exposure profile using the residence address as a surrogate for exposure assessment. This project is one of the largest prospective studies to date investigating ambient air pollution exposure and breast cancer risk, and it should significantly contribute to increase current knowledge on the health effects of air pollution.

The XENAIR project has received a financing from the call "CANC'AIR"of the ARC foundation for cancer research in 2015, covering a 4-year period (2016-2020).

1.4 Air pollution modeling for epidemiological studies

One of the most critical issues of epidemiological studies associating air pollution to health effects is the evaluation of the population exposure to a given pollutant. An increasingly quantity of studies has been carried out to assess exposure at individual level, showing that very different approaches are feasible.

The first dominant approach was the application of an exposure value at a central site to the entire population of the study domain, assuming that pollutants are homogeneously distributed within large urban areas. Several studies suggested that greater variations are present at intra-urban level and that this method may lead to the misclassification of the personal exposure and significantly alter the health outcomes in the epidemiological results [Briggs et al., 2000]. Great attention is then given to modeling approaches that describe the spatial and temporal variability of a certain pollutant specie within a certain domain, which results in air pollution maps with a given resolution. Models outputs can predict future exposure or reconstruct historical exposure [Zou et al., 2009]. Furthermore, the degree of complexity (pollutants considered, spatial resolution) can be defined in function of the assessment needs and of the available input data. The nature of exposure modeling can be both statistical and deterministic and there is an increasingly diffused tendency to couple models with Geographical Information System (GIS), allowing to manage both the pollutant concentration data and the distribution of the epidemiological cohort's subjects.

Geographic Information Systems in Environmental Epidemiology

Advances in geographic information systems technology facilitate epidemiologists to study associations between environmental exposure and the spatial distribution of a certain disease. A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze,

manage, and present spatial or geographic data. GIS applications are tools that allow to create maps, do spatial analyze and edit data. In the context of air pollution impact assessment on human health, due to the high spatial variability of air pollutants, one of the key aspects is to have a high precision in term of subject's geolocalisation. Cohort members are usually geolocated through the geocoding process, which is the turning of textual address data into geographic representations, estimating its location coordinates. The validity of epidemiological studies on air pollution impact strictly depends both on the proportion of addresses that can be geocoded and on the positional accuracy of the geocoding process [Bonner et al., 2003]. Location data for the study population are corresponding to the actual residence of the cohort's subjects or, alternatively, to a set of geopolitical units (addresses, census blocks, neighborhood centroids) [Nuckols John R. et al., 2004].

Ward et al. [2005] compared Global Positioning System (GPS) measurements with locations obtained by geocoding subjects' addresses with the GIS and concluded that, despite having some inherent problems, most of the addresses located in towns can be geocoded without large errors. Bonner et al. [2003] conducted a similar study in Western New York State indicating a median distance between GPS and GIS of 38 meters and concluding that, for the most part, geocoding of addresses is a very accurate process. On the contrary, a case-study in Orange Country (Florida) investigating the geocoding quality in exposure for children living near high traffic roads suggested that typical street geocoding is insufficient for fine scale analysis [Zandbergen, 2007]. However, the recent improvements of the GIS software have permitted to increase the accuracy and the completeness of located addresses. A study recently conducted into the Auvergne-Rhône-Alpes region (France) demonstrate that geocoded addresses, even though not initially designed to be used for environmental exposure assessment, could be feasible in epidemiological studies [Faure et al., 2017].

Proximity and spatial interpolation models

Proximity modeling is a very simple approach for the exposure estimation to air pollution. A proximity model measures the distance of a receptor to a pollution source assuming that the exposure at a location nearer to an emission source is greater than at further locations, creating a proxy variable which is proportional to the exposure level of the population's members. Pless-Mulloli et al. [1998] investigated the occurrence of lung cancer among people that lives close to industries in Teesside and Sunderland (UK) categorizing subjects in three "zones"(near, intermediate, farther) in function of their distance to industrial areas. Other studies were performed considering the proximity to incinerators, hazardous waste sites or heavy-metals-emitting industries. Residential proximity to roads is the most widespread surrogate variable in epidemiological studies, as urban exposure to air pollution is mainly dominated by traffic emissions [Colvile et al., 2001]. Evidences that proximity to traffic could be a valuable proxy variable are diffused in literature. For example, Miyake et al. [2002] correlated distance from major roads with a series of health effects on Japanese adolescents. A similar study was published by Dadvand et al. [2014]: it associated the residential proximity to roads with term Low Birth Weight in Barcelona, observing that living within 200 m of major roads increase the term LBW risk of about 46%. A study in England and Wales investigated the association between air pollution and stroke mortality, adopting the distance from main roads as a proxy variable and observing that around 990 stroke deaths per year would have been attributable to road traffic pollution [Maheswaran and Elliott, 2003].

Another simple GIS-based approach to exposure assessment are spatial interpolation models. These methods estimate the value at a given location as a function of the values measured at surrounding monitoring stations. Spatial interpolation models are quite diffused because of their capacity to be adapted to each situation in function of available data or of the complexity which is required. Interpolation methods occupy a very widespread range of modeling approaches, that go from a very simple

Nearest-Air Quality Monitoring Station (AQMS) assessment to the kriging process.

Attention is given in finding whether a model simply based on the nearest air quality monitoring station measure could be a feasible and sufficiently accurate modeling approach for air pollution exposure assessment in epidemiological studies. This is of course justified by the fact that such a model is extremely easy to manage, requiring almost no input data. Since air pollution concentration are often measured on quite regular basis in many cities and data and statistics are often made available by the public authorities, this method is applicable to retrospective epidemiological studies in a very simple way. Nonetheless, results are controversial: Nearest-AQMS approaches are in fact extremely sensitive to the spatial resolution of the monitor network, and a low number of monitors within a certain domain could lead to great misclassification of the subjects exposure. Nerriere et al. [2005] conducted a study in 2005 comparing personal exposure data (taken by the subjects through samplers installed in rucksacks) and data provided by fixed monitors and concluded that some caution is needed in using the latter method. The main issue is related with the capability of the methods basing on monitoring stations datasets to capture spatial variability between subjects, since most of times AQMS measurements are representative only of pollution levels in the immediate proximity of the stations [Lebret et al., 2000].

More complex spatial interpolation method are those implying the Inverse Distance Weighting (IDW), that calculate the value at an unknown location as a weighted average of the measures at the surroundings monitoring stations, therefore assuming that the exposure value estimated is more influenced by the close measurements than the distant ones. Hoek et al. [2002] applied a model considering both the inverse-distance-weighted interpolation method and the proximity to roads to estimate concentrations of black smoke and nitrogen dioxide within the Netherlands. They observed some associations between air pollution and mortality due to cardiopulmonary diseases.

A more accurate weighted interpolation of measurements at surround-

ing monitors is named kriging. Kriging is a technique that assigns a certain weight at each concentration by maximizing the correlation among the measurement. It generates both estimates and standard errors that quantify the degree of uncertainty of the model. Künzli et al. [2005] published in 2005 a study associating ambient air pollution and atherosclerosis in Los Angeles and applied for exposure assignment a combination of universal kriging model with a multiquadric radial basis function model. This study represented the first epidemiological evidence of an association between atherosclerosis and air pollution.

Land Use Regression Models

More recently developed models called Land Use Regression Models (LUR) have combined proximity measurements with geographical factors (road type, land use, traffic variables), leading to the development of increasingly complete approaches that estimate the exposure level as function of the characteristics of the surrounding environment. The development of a LUR is made through the construction of multiple regression equations describing the relationship between the measured value at the monitor and a series of selected prediction variables. Typical variables considered are both related to the proximity of a pollution source and to other environmental factors that could be related with air pollution. Ryan and LeMasters [2007] identified typical classes and definitions of common geographic variables frequently included in land use regression models (table 1.1).

Many times, variables are chosen simply basing on the availability of data: for example, Briggs et al. developed a LUR model for Prague including the traffic as a predictor variable, while the model for Amsterdam refers on road length because no data were available about traffic [Briggs et al., 1997], [Briggs et al., 2000].

Class	Variable used	Variable definition
Road type	Road type 1	Road serving >25000 people
	Road type 2	Road serving 5000 - 25000 people
	Road type 3	Road serving 1000 - 5000 people
	Highways	Undefined
	Major roads	Undefined
	Major roads	Average daily traffic count >50,000 people
	High traffic roads	Road serving 10000 - 25000 people
	Minor road	Undefined
	Bus route	Public transportation route
Traffic count	Weighted traffic volume	15 * (Traffic volume <40 m) + (Traffic volume 40–300 m)
	Traffic volume	Traffic volume (1000 vehicle km hr-1)
	Traffic count on nearest highway	Undefined
	Average daily traffic count	Average number of cars traveling in both directions/weekday (vehicle-km/hr)
	Traffic intensity	Vehicles/day
	Heavy vehicle traffic intensity	Heavy traffic/day
	Average daily truck count	Average number of trucks traveling in both directions
Elevation	Altitude	Meters above sea level
Land cover	Land cover factor	Weighted sum of the areas of industrial and high density residential land
	Land cover	Area of built up land
	Industrial use land	Area of land designated for industrial use
	Open space land	Area of land designated for industrial use
	Commercia use land	Area of land designated for commercial use
	Government/industry land	Area of land designated for government or industrial use
	Household density	Number of houses in area
	Population density	Population in area
	Land use	Area covered by industry, heavy industry, multi-family residential housing
	Distance to coast	Distance to coast

Table 1.1: Classes and definitions of common geographic variables included in land use regression models [Ryan and LeMasters, 2007]

The classic equation provided by an LUR model to describe the pollution concentration is the following:

$$C = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_n x_n$$

where x_i are the different predictors and α_i are the coefficients resulting from the multivariate linear regression. The objective is to adjust the type and number of parameters and the coefficients to minimize the bias and increase the correlation with measurements.

After a LUR equation is formulated, a crucial step of its developing is the validation process, that consist in testing the model performance within the domain re-running the models after some monitors are removed. This step is called "cross-validation": despite there is not a universal procedure to conduct a cross-validation on LUR models and different studies often propose different methods, the Leave-One-Out-Cross-Validation (LOOCV) is one of the most diffused in literature [Wang et al., 2012], [Johnson et al., 2010].

LUR models have become quite a good alternative for air pollution exposure assessment in epidemiological research, being a very cost-effective method to explain the spatial variation in air pollution [Marshall et al., 2008]. On the contrary, one of the limitations that are often observed in LUR models is the fact of being quite site-specific. Moving between areas with different land use type and topology reveals the necessity to calibrate the model with local parameters, depending also on data availability (and on data quality of course) at the different locations. Other limitations are represented by the fact that they usually produce annual averaged (or biennial) estimation, while deterministic models can provide hourly concentration values. Furthermore, the development of a LUR models strictly needs a homogeneous distribution of measurement stations within the considered domain.

The European Study of Cohorts for Air Pollution Effects (ESCAPE) presented in 2013 a study describing a standardized way for LUR models developing applied to 36 study areas in Europe [Beelen et al., 2013]. The R^2 calculated for the models ranged from 55% to 92% for NO_2, with an aver-

age number of included predictors of 4 (all models included at least one traffic-related variable). Since the increasingly spatial resolution of GIS has considerably improved the precision and availability of traffic intensity data linked to digital road networks, a future improvement in land use regression models' performances is surely expected. Liu et al. developed in 2019 a land use regression model for the city of Xi'an in China, resulting in a 5 predictors model showing a $R^2 > 0.85$ [Liu et al., 2019].

Epidemiological studies using land use regression models for exposure assessment are increasingly diffused in literature. For example, Coogan et al. researched in 2016 a correlation between long term exposure to NO_2 and diabetes incidence, using both a dispersion model and a land use regression model to estimate concentration levels at residence address [Coogan et al., 2016]. Forastiere et al. investigated in 2019 the association with mortality of annual average air pollution exposure given by two different LUR models in Rome, a Europe-wide LUR and a local one [Forastiere et al., 2019]. They observed significant hazard ratios using both models for $PM_{2.5}$ and NO_2.

Dispersion Modeling

On the contrary to land use regression modeling that use a stochastic approach, dispersion models are the result of a deterministic process. Dispersion models simulate the physical and chemical processes of the dispersion and transformation of atmospheric pollutants so that they predict their concentration variability in space and time. They require both emission data and the basic meteorology, along with a simplified description of the domain geometry. Emission data can include both stationary and mobile sources: the first being local pollution sources (industries, waste sites, home heating), the seconds mainly related to traffic (usually estimated by road type, traffic flow, vehicle type). An important input data is also the ambient background concentration [Tchepel et al., 2010].

Common dispersion modeling methodologies are box models (where the domain is considered as a box in which pollutant are emitted and un-

dergo chemical and physical processes), lagrangian and eulerian models (define a region of air containing an initial pollutant concentration and then follow its trajectory as it moves downwind) and computational fluid dynamic models (CFD, provide analysis of fluid flow based on conservation of mass and momentum by resolving Navier-Stokes equations in three dimensions) [Holmes and Morawska, 2006].

However, dispersion models vary depending on the mathematics used the development, and the most commonly used are the Gaussian-based ones. These models are based on the fact that the time averaged pollutant concentration downwind from a source can be modeled using a normal (or Gaussian) distribution curve. The basic Gaussian dispersion model applies to a single punctual source (figure 1.3), but it can be modified to account for line sources o area sources [Gilbert and Wendell, 2014].

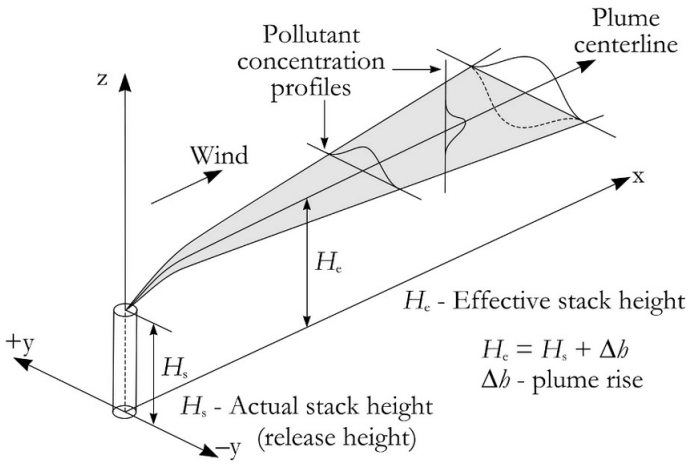

Figure 1.3: Plume dispersion coordinate system, showing Gaussian distributions in the horizontal and vertical directions [Gilbert and Wendell, 2014]

The normal distribution of the plume is modified at greater distances due to the effects of turbulent reflection from the surface of the earth and at the boundary layer when the mixing height is low. The width of the plume is determined by coefficents ($\sigma_{y,z}$) defined by stability classes of the atmosphere.

Lots of Gaussian-based dispersion models have been developed by public authorities: the California Departement of Transportation developed a Line Source Dispersion Model named CALINE to predict concentration of CO, NO_2 and PM near highways and arterial streets [Benson, 1988]. American Meteorological Society and the Environmental Protection Agency proposed a near field steady state model for particle dispersion named AERMOD in 2005 [Cimorelli et al., 2005], that was lately expanded to gas phase pollutants.

One of the major challenges for dispersion modeling development is the description of pollutant behavior in street canyons, which is a term frequently used for urban streets flanked by buildings on both sides. Lots of dispersion models were specially developed or simply used to street network applications, as reviewed by Vardoulakis et al. [2003]. SIRANE is an air pollution dispersion model for an urban environment: it decomposes the domain in a urban canopy (where pollutant flows are simulated into a simplified geometry of the street network) and the external atmosphere, where street intersection and stationary sources are modeled as Gaussian plumes [Soulhac et al., 2011]. Further details on SIRANE are given in section 2.3.

Comparison between LUR models and Dispersion Models

High resolution concentration maps over large periods of time have now become crucial in environmental epidemiology to realize precise risk assessments, since measurement of individual participants are often impossible (especially for retrospective studies). As explained in the previous sections, dispersion modeling and land use regression (LUR) modeling are two of the approaches that are currently widely used for small-scale spatial variations in air pollution concentrations.

Dispersion models are very accurate but cannot cover large areas, being rather specifically applied for urban-scale simulations. LUR models are increasingly used since they allow to simulate pollutant concentrations over countries or even continents [Beelen et al., 2013], taking into account

that national-scale simulations unavoidably imply a loss of information at local scale. Moreover, since those two methods are conceptually opposite (while LUR models are empirical, statistical models, dispersion models are based upon physical principles and their mathematical description) the comparison between these two approaches is extremely useful to observe their relative performances in estimating air pollution concentrations at small-scale within a urban domain, assigning exposure values to cohort's members. The purpose of studies comparing LUR and dispersion models is to quantify these differences and their relative importance, mainly focusing on assessing whether and how much they have an impact on the results of epidemiological studies.

Since now, only a few studies compared the performances of LUR modeling and dispersion modeling in estimating air pollution concentrations.

Cyrys et al. [2005] used both a stochastic model and a dispersion model (IMMIS$^{net/em}$) to predict NO_2 and PM_{10} concentrations in Munich, Germany, at 1669 addresses of the participants of two ongoing birth cohort studies. IMMIS$^{net/em}$ describes the dilution and transport of pollutants from point, line, and area sources as a stationary process, using a Gaussian normal distribution. The results showed a strong correlation between stochastic- and dispersion- modeled concentrations for both pollutants.

Marshall et al. [2008] compared three approaches for estimating within-urban variability in ambient concentrations of NO, NO_2, CO, O_3 at 56099 postal codes in Vancouver (Canada): a GIS-based model for spatial interpolation of monitoring data, a LUR model and an eulerian grid model (CMAQ). In general, the three approaches reflected different spatial scales: urban-scale variations for interpolated ambient monitoring data and the dispersion model, neighborhood-scale variations for LUR. Differences in means and standard deviations among the methods were modest, even if LUR exhibited higher spatial resolution than the other methods.

Beelen et al. [2010] compared the performances of a LUR model and a dispersion model (URBIS Information System) in estimating NO_2 concentrations in a Dutch urban area (Rijmond area, corresponding to Rotterdam and surroundings). The regional background was obtained by interpola-

tion of regional measurements and concentration data were estimated for 70000 centroids on a regular grid of 100x100m. A moderate agreement was found (Pearson's r = 0.55) especially for the central part of the exposure values' distribution: the main differences were observed to be due to the land use category *industry* into the LUR predictors and to the different treatment of the NO-NO_2 conversion.

de Hoogh et al. [2014] explored the differences between LUR and Dispersion Models estimates for NO_2, PM_{10} and $PM_{2.5}$ within the European Study of Cohorts for Air Pollution (ESCAPE project), developing LUR models basing on a standardized methodology. 13 areas were involved for NO_2, 7 PM_{10} and 4 for $PM_{2.5}$: LUR and dispersion model estimates correlated on average well for NO_2, with median Pearson's r and Spearman's ρ respectively equal to 0.75 and 0.77 (this implies that both methods may be useful for epidemiological studies of small-scale variations of outdoor combustion related air pollution, typically from road traffic) but only moderately for PM, with large variability across different areas.

Wang Meng et al. [2015] compared the agreement between long-term air pollution exposure estimates for NO_2, PM_{10}, $PM_{2.5}$ and soot based on dispersion modeling and LUR modeling. Also, they evaluate whether associations between long-term air pollution exposures and lung function in children differ depending on the exposure modeling approach used. Participants were included from the Dutch PIAMA (Prevention and Incidence of Asthma and Mite Allergy) birth cohort study, counting 3963 newborns. Overall, the LUR model predictions correlated well with the estimates of the dispersion models for all the pollutants. Also, in this study, a better agreement was observed for NO_2 (r = 0.86 for NO_2, 0.57 for PM_{10}).

Hennig et al. [2016] compared a LUR and a Dispersion and Chemistry Transport Model (DCTM) in the Ruhr area, Germany, using 4809 residences' coordinates. The correlation they observed was weak to moderate, attributed to the fact that LUR and DCTM models do not represent identical aspects of air pollution: while DCTM represents an area average similar to urban background concentrations, the ESCAPE-LUR was designed to predominantly estimate variability in local traffic-related air

pollution.

Objectives of the study

It is certainly important to carry out further analyses in comparing dispersion and land use regression modeling approaches, with the aim of better understanding how the choice of the model can impact the estimation of the risk of breast cancer occurrence related with high air pollution exposure, and so affect public decisions about healthcare strategies.

This work is part of the XENAIR project and has the objective of comparing the results of a national Land Use Regression model with those of a dispersion model (SIRANE). The focus will be on the loss of information when passing from a deterministic model providing spatially refined estimated concentrations in a relatively small domain to a stochastic national approach (table 1.2):

model	*Spatial resolution*	*Temporal resolution*	*Domain dimension*
SIRANE	10 meters	Hourly time-step	Lyon metropolitan area
LUR	50 meters	Yearly average	France

Table 1.2: Spatial and temporal resolution of SIRANE and the LUR model simulations' results

A retrospective comparison have been made between annual average exposure values estimated in 2010 and 2000 for a real case-control cohort (E3N, [Clavel-Chapelon and E3N Study Group, 2015]), further investigating if the loss of information could be attributable to specific land use types or socioeconomic factors. The pollutant considered were nitrogen dioxide (NO_2), ozone (O_3) and particulate matter with aerodynamic diameter smaller than 10μm (PM_{10}). Proximity and interpolation models were also involved into the comparison for the year 2010. Additionally, to quantify the impact of this difference on epidemiological results, the two models underwent a comparison for the calculus of typical epidemiological indicators (odds ratio).

2 Methods

2.1 *Study area*

Figure 2.1: Lyon position in Europe

Lyon is the third largest city and second-largest urban area of France and the capital of the Auvergne-Rhône-Alpes region, located in the country's east-central part (figure 2.1). In 2017, Lyon had a population of 516,092 habitants (2,326,223 for the metropolitan area). The climate is classified as semi-continental with mediterranean influences. Lyon's geog-

raphy is dominated by the Rhône and Saône rivers that converge to the south of the city center forming a peninsula, and two large hills are situated at the north and at the west of the downtown. The study domain is a rectangular area of around 1190 km^2 that extends in latitude from Givors up to the north of Lyon and in longitude from the countryside at the east of the city to the Saint-Éxupéry airport. The domain includes 143 municipalities and occupies three French departments: *Rhône-et-Loire*, *Isère* and *Loire*. Outputs of various modeling approaches have been applied within the domain showed in figure 2.2b, that is the intersection of all the models' domains.

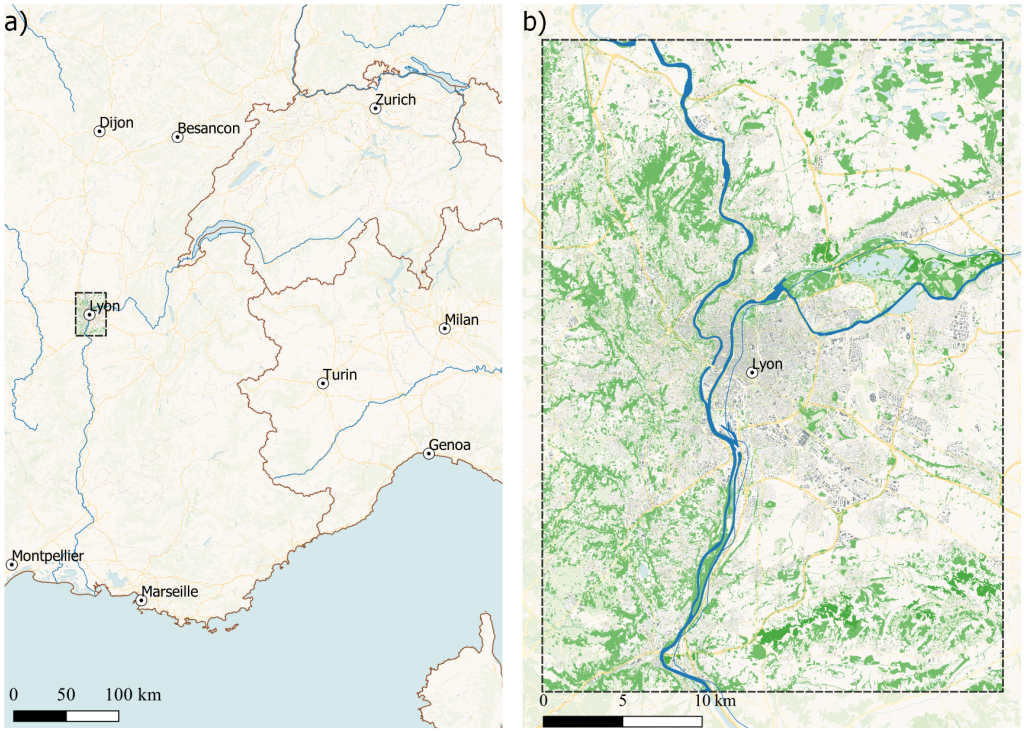

Figure 2.2: Study domain

The domain presents both rural and urban areas: following the CORINE Land Cover protocol [Büttner, 2014], there are artificial surfaces for the 38.46 % of the surface. Others are agricultural areas (49.65 %), forest and seminatural areas (9.31 %), water bodies (2.42 %) and wetlands (0.07 %).

2.2 Model CHIMERE

The model CHIMERE, in addition to being included in the comparison for the year 2010, is of crucial importance for the operation of the LUR and SIRANE models, as explained in section 2.3 and 2.4. CHIMERE is an atmospheric pollution model, dedicated to studies about events at regional scale. Those are resulting of high emissions (both anthropogenic and natural), stagnant meteorological condition but also of the kinetics and efficiency of the chemistry and the deposition. More specifically, CHIMERE is an Eulerian off-line chemistry-transport model (CTM). As input data, the model considers the primary pollutant emissions, the meteorological fields and the chemical boundary conditions. The domain can vary from continental to local (from 1 km to 1 degree resolution).

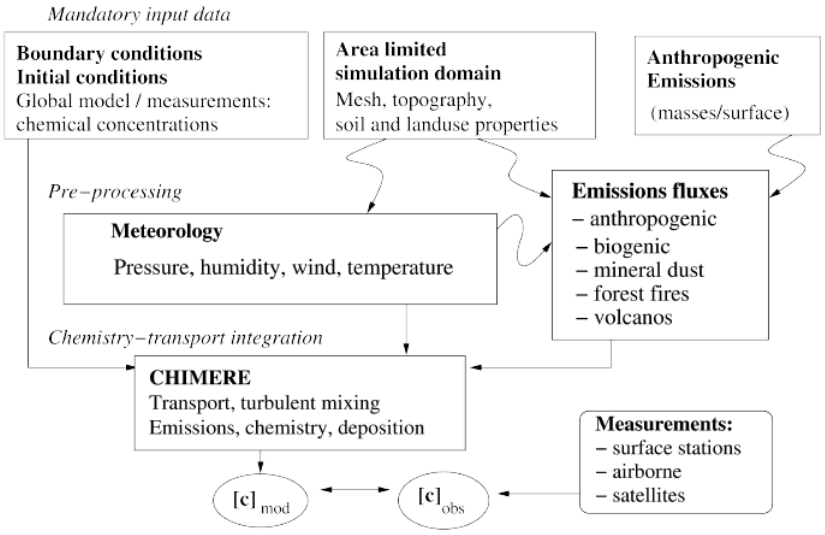

Figure 2.3: General principle of a chemistry-transport model; $[c]_{mod}$ and $[c]_{obs}$ are the modelled and the observed chemical concentrations fields, respectively

Atmospheric concentration fields of tens of gaseous and particulate pollutant species are the outputs of the simulation, and the processes that mainly affect the results are the emissions, the transport phenomena,

the chemical reactions and the deposition. Figure 2.3 presents a general principle of CTM. The first version of the model was released in 1997 including only gaseous species and covering the Paris area [Vautard et al., 2001]. Now the CHIMERE model is considered a state-of-the-art model [Menut et al., 2013], being involved in numerous studies all over the world ([Schaap et al., 2007], [Zyryanov et al., 2012], [Hodzic et al., 2009]). In this study, CHIMERE simulation for the year 2010 were used for NO_2, PM_{10} and O_3 exposure estimation, focusing on annual average values.

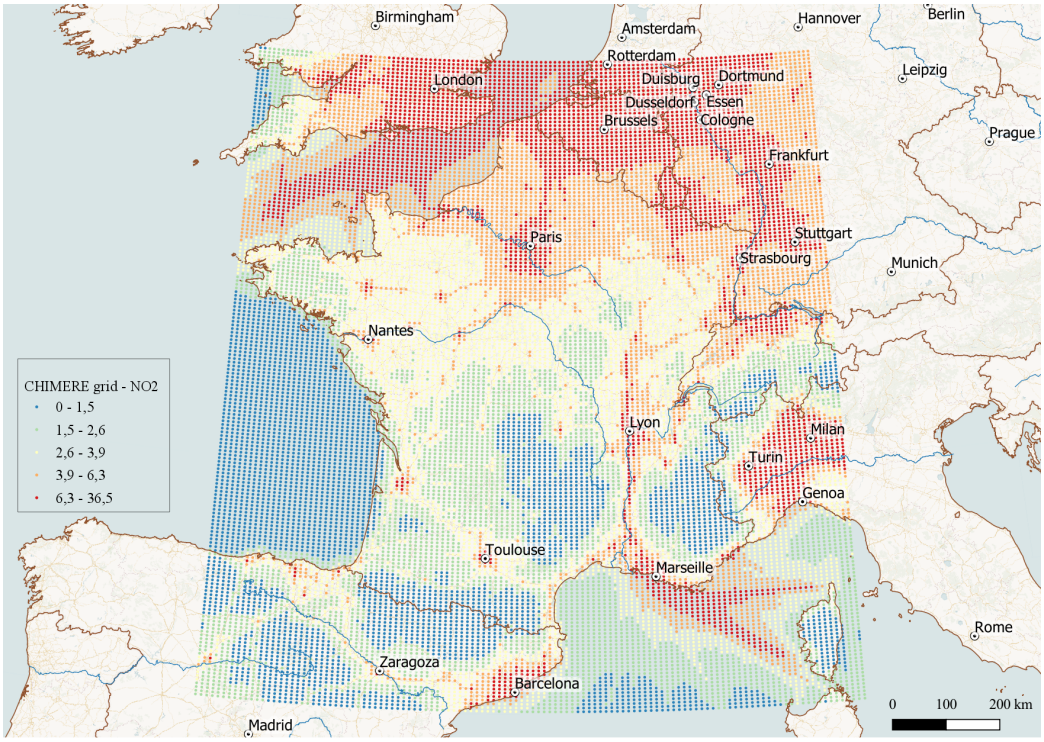

Figure 2.4: CHIMERE grid within France; legend values are in $\mu g/m^3$

Meteorological and flux data were given from the European Monitoring and Evaluation Program (EMEP) and concentrations were provided with a spatial resolution of approximately 7 km x 7 km, which is a quite fine grid for a CTM model given the overall extension of the domain. Schaap et al. performed in 2015 a study to investigate the impact of using finer grids resolution in CTM, comparing four models including CHIMERE [Schaap et al., 2015]. They observed that decreasing the grid

scale is very helpful for underlying the "urban signal", namely the difference between high emission areas and their surroundings, especially for PM_{10} and NO_2. On the contrary, ozone concentrations are less affected by model resolution [Queen and Zhang, 2008]. CHIMERE outputs was implemented into the GIS is in form of a punctual layer, with the points set as a grid all over the domain, as displayed in figure 2.4 for France and in figure 2.5 for the study domain.

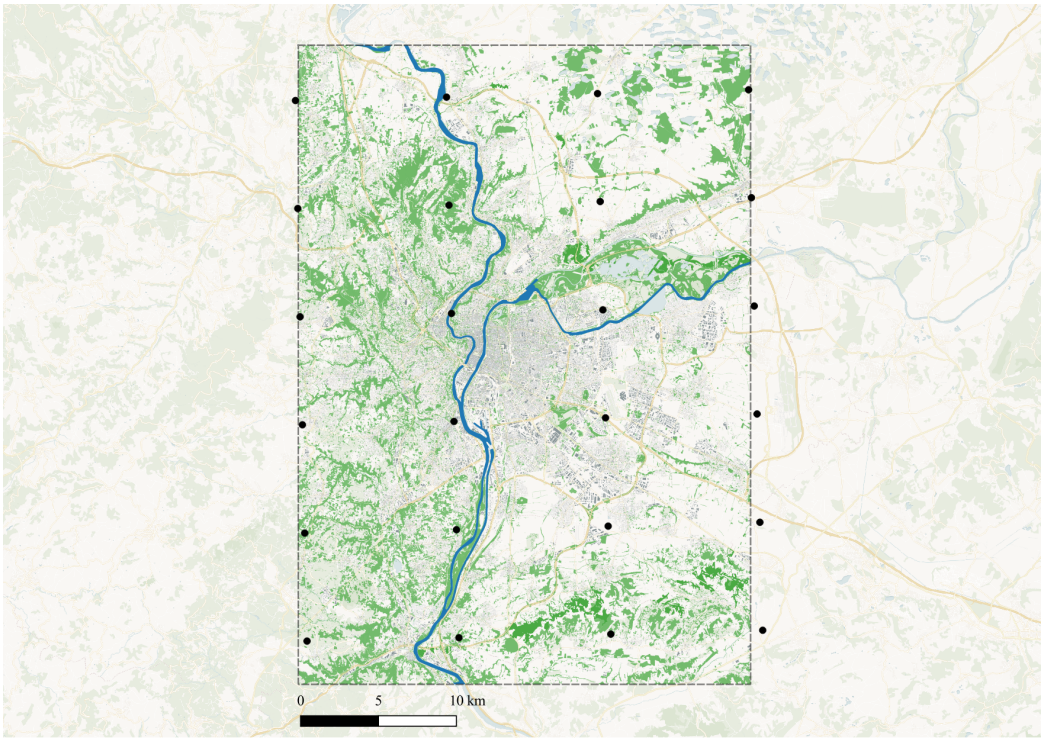

Figure 2.5: CHIMERE grid within the domain

The exposure assignment method starting from the CHIMERE grid within the domain was computed with the GIS and is explained in section 2.6

Even though the grid employed in this study has a quite fine resolution for a CTM model, it belongs to a simulation made at a regional scale and logically shows a weaker resolution level compared with others fine scale models that will be presented.

2.3 Model SIRANE

SIRANE is an urban dispersion model, developed by the *Laboratoire de Mécanique des Fluides et Acoustiques* of the *École Centrale de Lyon* and presented in 2011 [Soulhac et al., 2011]. The model is based on a decomposition of the domain in two parts: the urban canopy and the external atmosphere, managed by two independent modules.

Pollutant transfers within and across those modules are parametrized, as a function of meteorological data (wind speed and direction, temperature, cloud cover and precipitation intensity). Pollutant dispersion and deposition (both dry and wet) is simulated with an hourly time-step. Source typologies considered in SIRANE are both industrial emissions, represented as elevated point sources, and traffic emissions, as line sources distributed on a road network. Miscellaneous diffuse sources (such as domestic heating) are also considered and represented as areal sources at ground level.

The model performs a simplified description of the urban geometry, where streets are modelled as a simplified network of connected segments which are represented by boxes, as showed in figure 2.6.

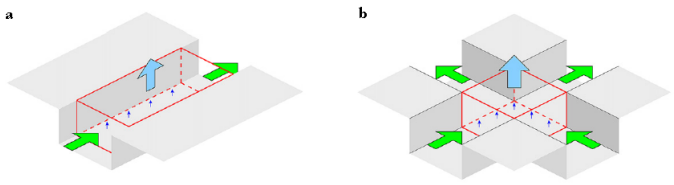

Figure 2.6: Simplification of urban geometry in SIRANE. a) Box model for each street with relative flux balance. b) Network of streets [Soulhac et al., 2011]

The mass transport simulation considers three mechanisms: a convective flux along the streets (due to the parallel component of the external wind speed, the green arrow in figure 2.6a), a turbulent transfer across the boundary urban canopy - external atmosphere (blue arrow in figure 2.6a)

and a convective transport at street intersection [Salizzoni et al., 2009], [Soulhac et al., 2009]. An important assumption is that the pollutant is assumed to be perfectly mixed inside each street segment.

In the external atmosphere, the flow is described by the Monin-Obukhov similarity theory [Pahlow et al., 2001]. As a roughness sub-layer is not considered above the urban canopy, the external flow is assumed to be uniform and the dispersion of the pollutant advected or diffused within the external atmosphere is described with a Gaussian plume model (figure 2.7).

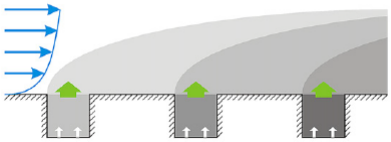

Figure 2.7: Gaussian plume modelling for pollutant transport above the urban canopy [Soulhac et al., 2011]

The model has been validated, comparing its results to field data measured within an urban district in Lyon, France [Soulhac et al., 2012]. A measurement campaign 15-day long conducted in the VI *arrondissement* (named LYON6) provided information about traffic fluxes and cars emissions, meteorological conditions, background pollution levels ad spatial variability of pollutant concentrations. The overall comparison between model predictions and field measurements was classified as 'good' following criteria from Chang and Hanna [2004]. The same result was obtained during another validation study over a whole urban agglomeration (Lyon) in the year 2008 for nitrogen dioxide [Soulhac et al., 2017].

One of the major problems for the modelling of pollutant concentrations at urban scale is to estimate a background concentration [Tchepel et al., 2010]. This concentration is associated to the contribution of all pollutant sources located outside the studied domain, in the way that the values predicted by the model exactly correspond to the excess above the

background ones. Possible approaches to define background concentrations for local scale models are to use monitoring air quality data or using simulation results from larger domain models [EPA, 2005]. In case of estimation via measurement stations, it is crucial that those stations are placed at the border of the domain, far away from traffic axes [Dėdelė and Miškinytė, 2015]. For the validation study over Lyon in 2008, the background concentration value was measured at the Saint-Éxupéry Airport, located at the east border of the domain (approximately 30 km from the city center) [Soulhac et al., 2017].

Models simulation realized for the XENAIR study

For the year 2010, simulation outputs of two different results of SIRANE were available: the "Saint-Éxupéry"(SE), in which the background concentrations included were the average values measured at the Saint-Éxupery airport, and the "Extraction"(EXT), that instead uses concentrations estimated by a CHIMERE simulation in correspondence of the same location.

Table 2.1 presents background values for the two results of SIRANE

	NO_2	O_3	PM_{10}
SIRANE Saint-Exupéry	14.52	11.91	25.06
SIRANE Extraction	15.90	9.51	15.91

Table 2.1: Averaged background concentration values for SIRANE results in 2010; all values are in $\mu g/m^3$

Figure from 2.8 represent the simulation results over the city of Lyon for NO_2 in EXTRACTION, while figure 2.9 show a zooming within the downtown and superposed with satellite images. All the results for both versions are presented in Appendix C.

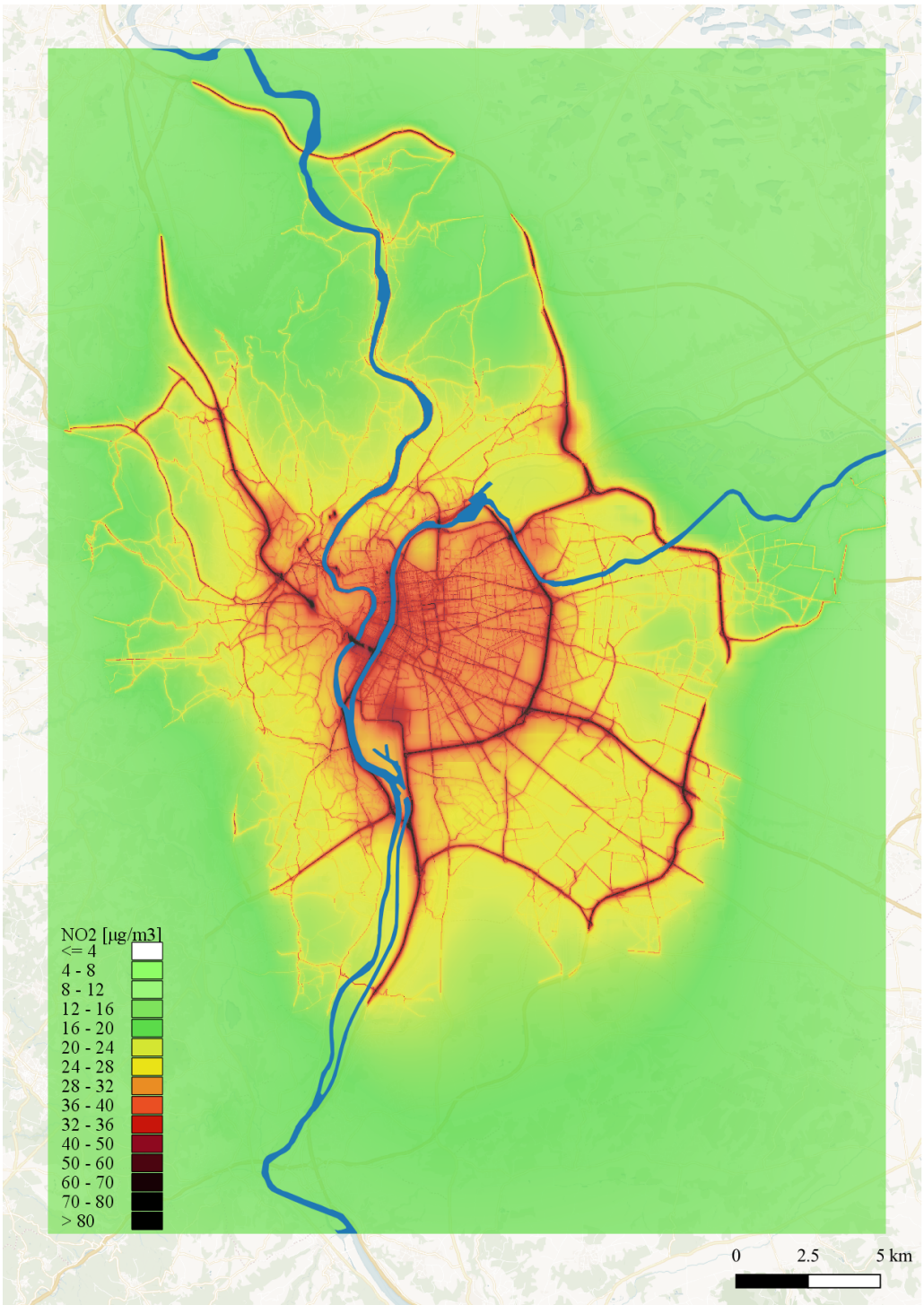

Figure 2.8: Result of the SIRANE EXT simulation in 2010 for NO_2

Figure 2.9: Result of SIRANE EXT simulations in 2010 for NO_2. zoom within the city center

The main difference between the two results of SIRANE are surely related to the PM_{10} concentration estimation, given the quite relevant difference in term of background concentration among them, almost equal to 10 $\mu g/m^3$ (25.06 vs 15.91 $\mu g/m^3$, see table 2.1).

Since a SIRANE output with measured background was not available for the year 2000, the choice was to include into this study the EXT one. This is justified by the fact that, since the XENAIR project involves retrospective studies with exposure simulation every 5 years from 1990 to 2010, comparisons between different periods need to be performed among models whose background values were defined using the same method. Figure 2.11 shows results for SIRANE EXT in 2000.

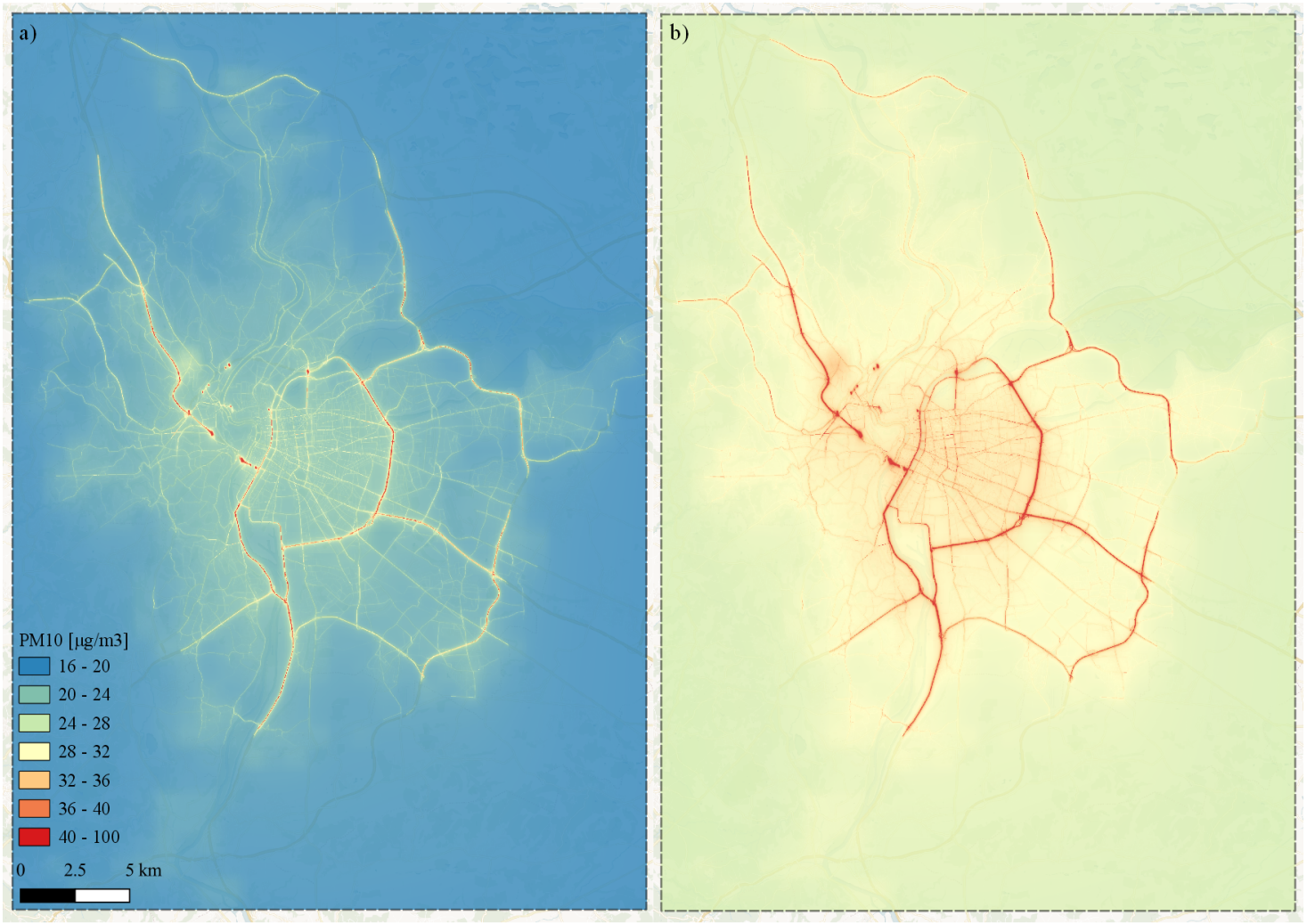

Figure 2.10: Result of SIRANE simulations in 2010 for PM_{10}; a) version EXTRACTION; b) version SAINT-ÉXUPERY

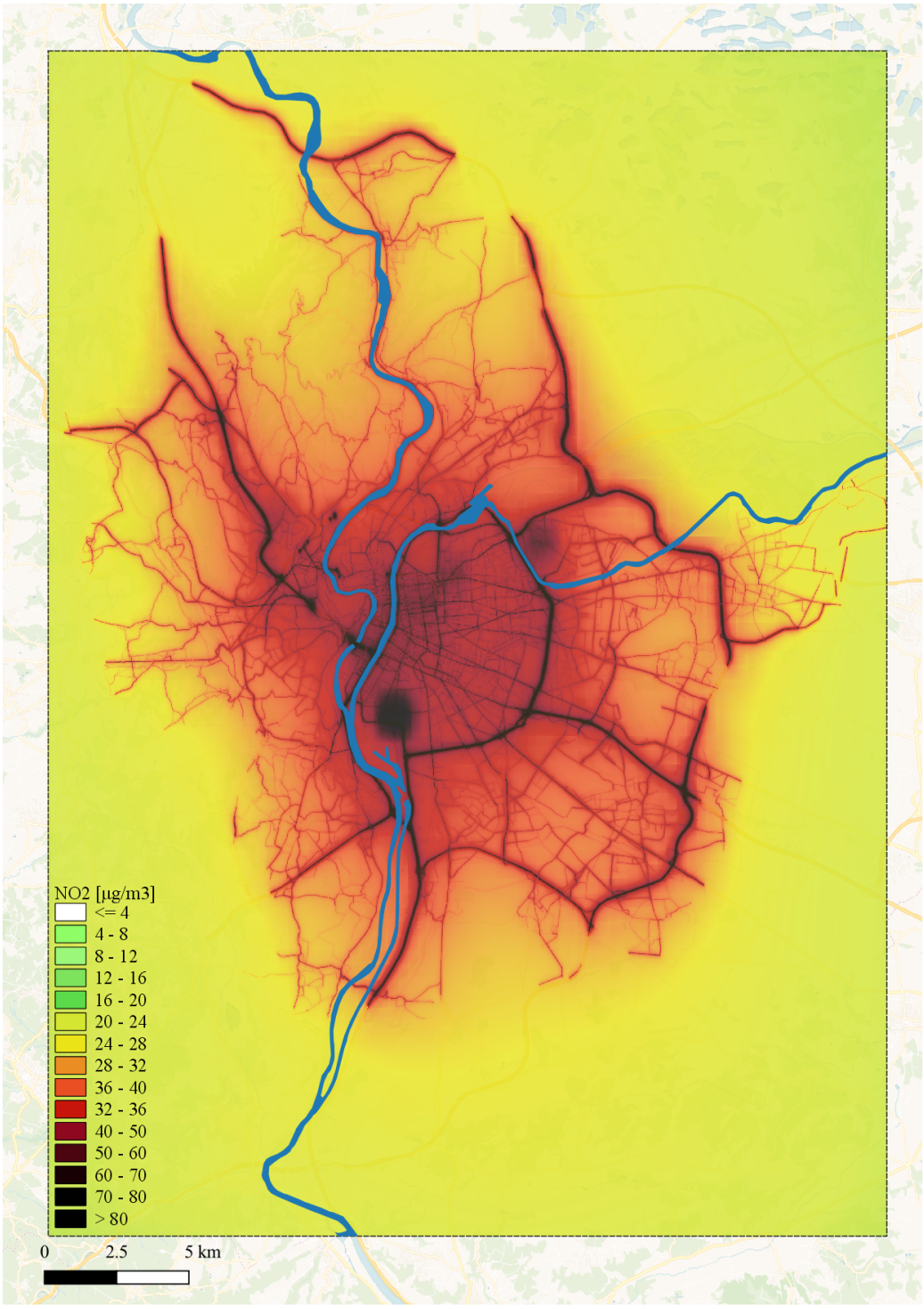

Figure 2.11: Result of SIRANE EXT simulation in 2000 for NO_2

2.4 Land Use Regression Model

The second model applied within the domain to explore spatial variability of NO_2, O_3 and PM_{10} was the land use regression model developed and applied under the XENAIR project. The model covers all the European area of the French territory (figure 2.12), estimating several pollutant concentrations with a 50x50 meters resolution.

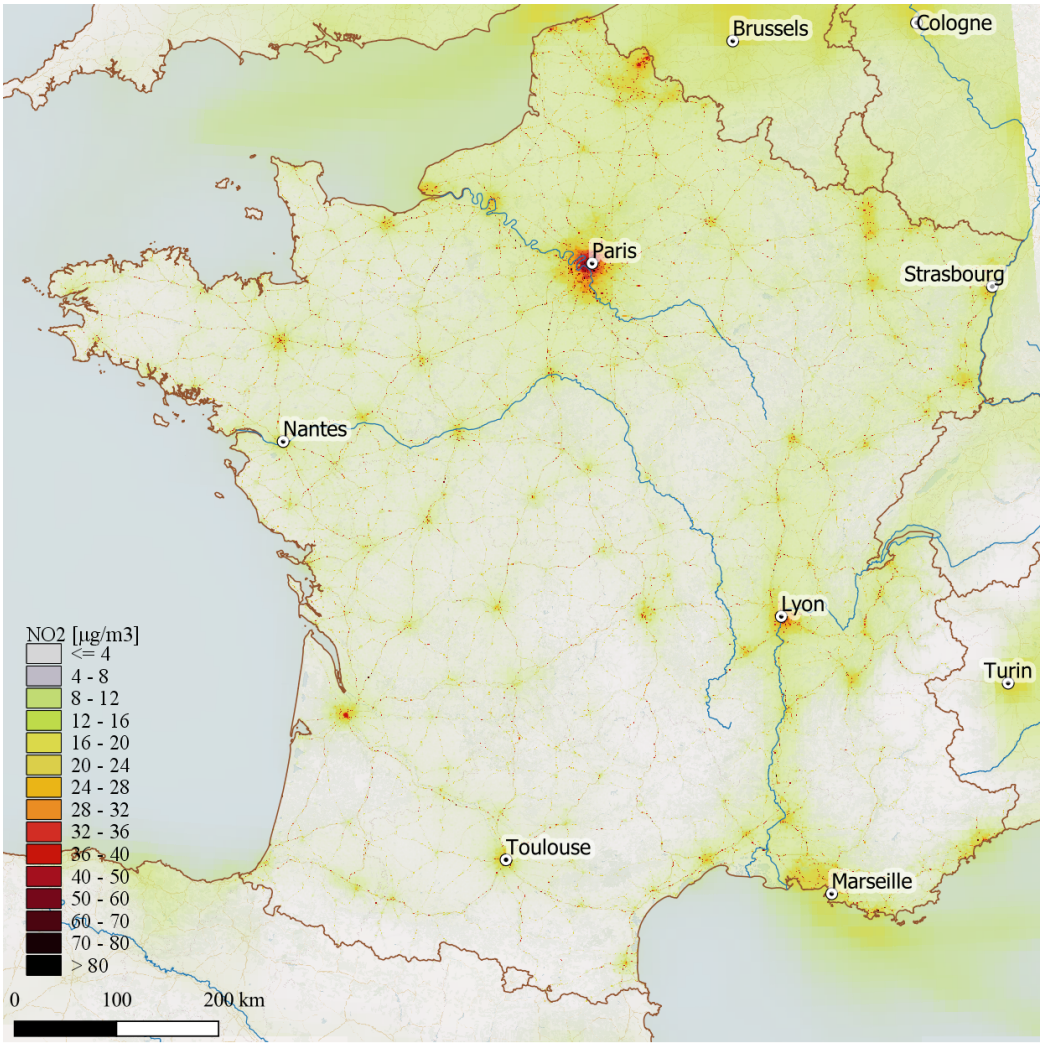

Figure 2.12: Result of the LUR for NO_2 in 2010 considering the whole simulation domain

Technically, the model can be classified as a "hybrid"model, because concentration values estimated by the CHIMERE model were involved into the predictors. This technique was also applied for a LUR model developing in the Ruhr area [Henning et al., 2018]. The model domain occupies an area slightly wider than the French territory (figure 2.12). The building of the model referred to the measured values given by AirBase, the air quality database maintained by the European Environmental Agency, using around 360 monitors all over France [https://www.eea.europa.eu /data-and-maps/data/airbase-the-european-air-quality-database].

Considering the LUR model developing procedure, variables must be chosen in order to minimize the difference between observed and predicted concentrations. This is usually done using statistical indicators as R^2 and RMSE. The procedure starts from a univariate regression analysis between the measured concentrations and all the potential variables. Then, a first predictor is defined, which is the variable giving the greatest R^2, and having previously defined its direction of effect (for example, positive for major road length).

The remaining variables are then added separately and the increase of the model accuracy (R^2, RMSE, Fractional Bias) is each time assessed: only variables leading to a R^2 increase of a minimum pre-defined value (usually 1%) are kept into the model. Finally, variables which had a low p-value are usually excluded.

The model is then validated through a variation of LOOCV (Leave-One-Out-Cross-Validation), consisting in re-applying the model versus the monitors that have been used to build it, each time leaving 20% of them, and assessing the average R^2 resulting from all the applications.

Average values for the year 2000 and 2010 were calculated for NO_2, O_3 and PM_{10}. Figure 2.13 shows the NO_2 distribution into the domain for 2010, while all other figures are presented in Appendix C.

Predictors resulting from the multiple regression performed for the XENAIR LUR model are presented in table 2.2, with relative resulting R^2:

Pollutant	Predictor	Buffer type (m)	Global R^2
PM_{10}	CTM MACC	Nearest point	0.59
	Major road length	50	
	High density urban	500	
	Agriculture and forest	10000	
NO_2	CTM MACC	Nearest point	0.67
	Road length	1000	
	Major road length	50	
	High density urban	500	
	Industry	10000	
O_3	CTM MACC	Nearest point	0.6
	Low density urban	3000	

Table 2.2: Predictors and R^2 values for the XENAIR LUR model

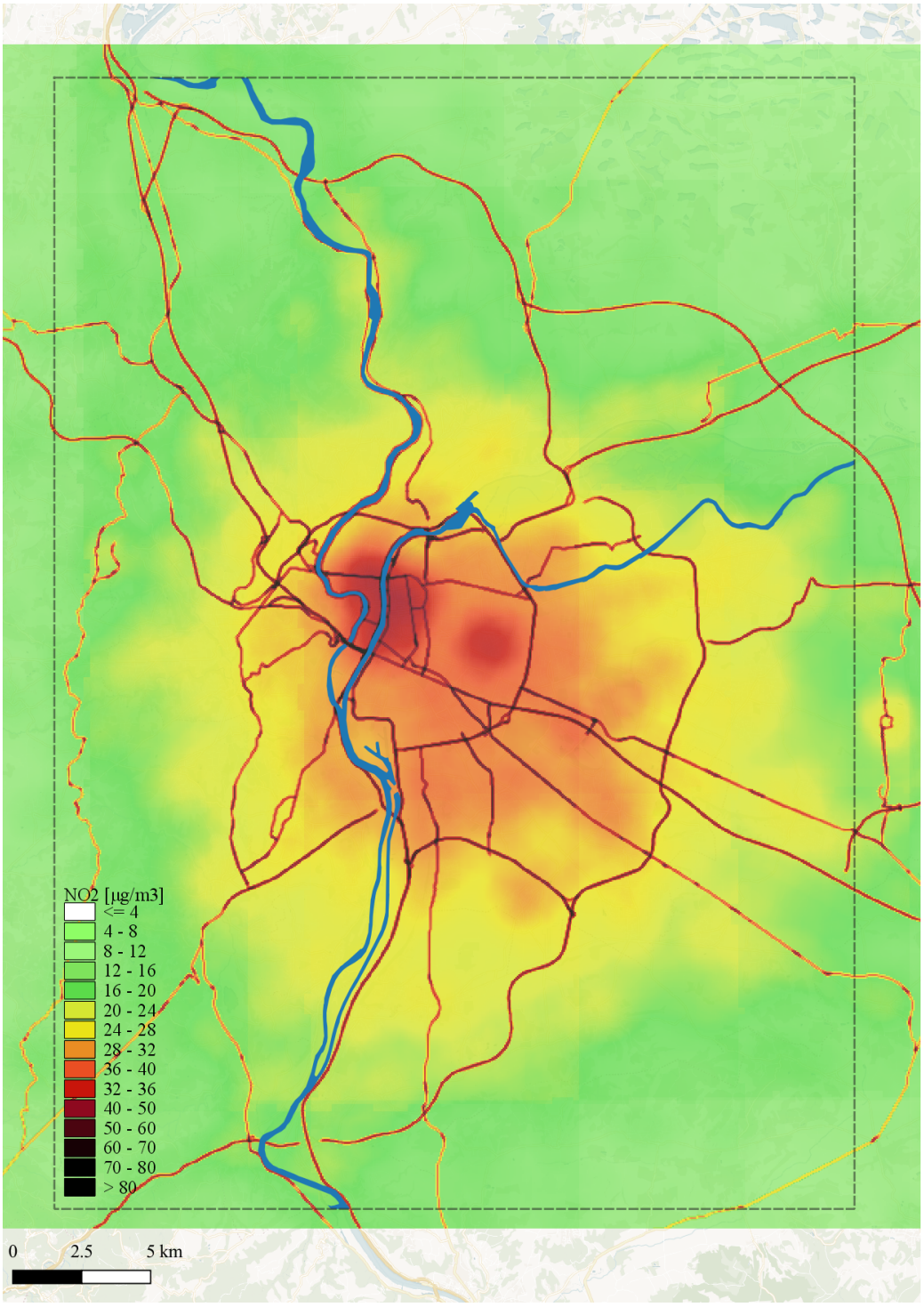

Figure 2.13: NO_2 LUR results for 2010 within the study domain

2.5 *Populations*

2.5.1 Overview

One of the study's objectives is to assess and compare different models' performances in term of estimating the exposure to air pollution for epidemiological studies, making necessary to refer to an epidemiological cohort or to equally representative surrogates. Different vector layers were implemented into the GIS referring to geocoded addresses, both representing real individuals addresses and randomly geocoded ones:

- One real population, composed by real subjects (members of the E3N epidemiological cohort);
- One virtual population obtained by a random selection between building's addresses of the city of Lyon;
- Two "semi-random"population, obtained by a random points creation within the domain in function of the population density;
- One random population, fully randomly created within the domain.

The choice to use also other populations in addition to the real one is justified by the need to verify that the results obtained by the comparisons between different models are not affected by the way the sampled values within the domain are chosen.

2.5.2 Real population

The real population has been built using the location of the members of the E3N Study Group resident into the domain boundaries. The E3N cohort (*Étude Épidémiologique auprès des femmes de la Mutuelle Générale de l'Éducation Nationale*) was initiated in 1990 to investigate the risk factors associated with cancer and other non-communicable diseases in women [Clavel-Chapelon and E3N Study Group, 2015]. Nearly 100000 women volunteered, required to fill questionnaires every 2-3 years and to submit

a signed consent from providing permission to obtain personal information (vital status, address, medical expense reimbursements from the insurance plan). The questionnaires are available at [http://WWW.e3n.fr/]. Several studies have been performed basing on this cohort, both for epidemiologic issues related to the exposure to air pollution ([Amadou et al., 2019] , [Danjou et al., 2015]) and for specific investigations in medical context (Fournier et al. researched in 2007 a possible relationship between the risk of breast cancer and different hormone replacing therapies [Fournier et al., 2008]).
In the XENAIR project, as said before, nearly 10000 women were involved in France for the year 2010. In figure 2.14, the E3N cohort is showed, while figure 2.15 presents the detail of the cohort members located within the study domain, the intersection resulting in 785 subjects.

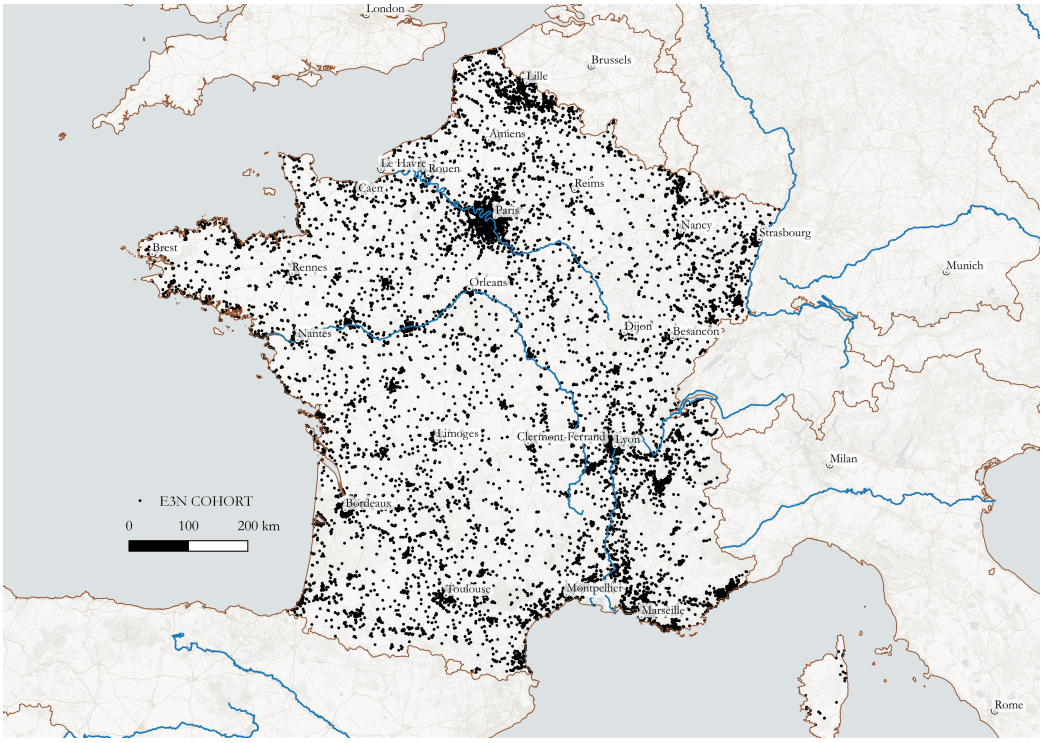

Figure 2.14: Georeference of all the E3N cohort members

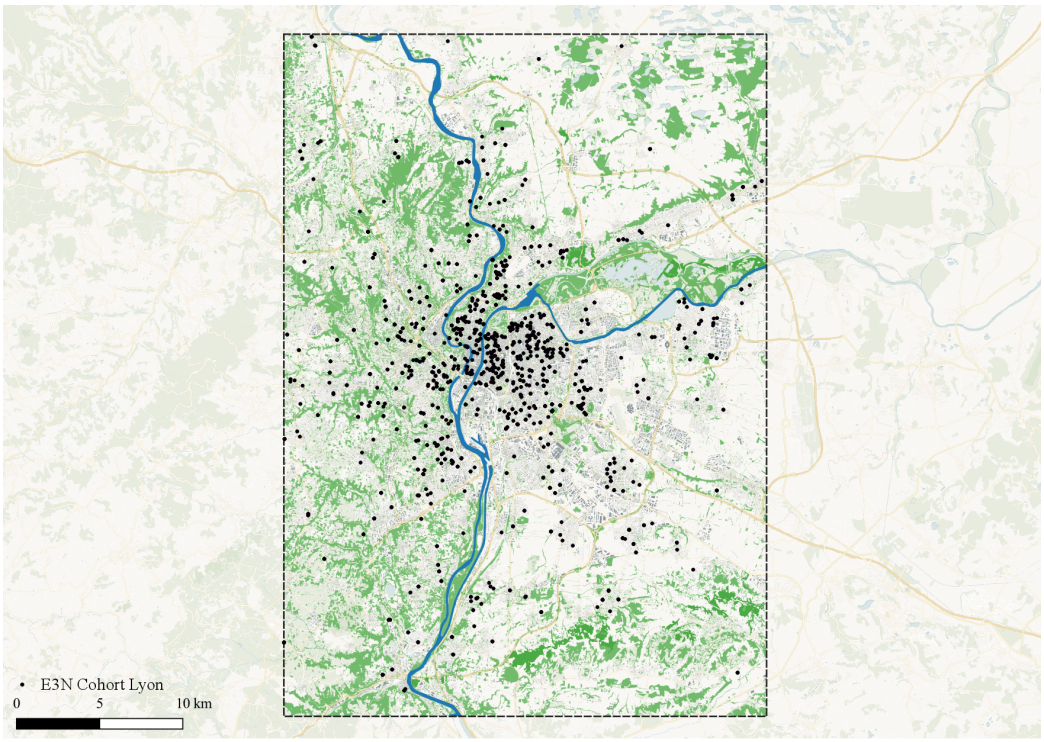

Figure 2.15: Georeference of the 785 E3N cohort members within the study domain

2.5.3 *"Points d'addressage"* population

The PA (*Point d'addressage*) populations have been created basing on the data describing all the buildings addresses into the *Métropole de Lyon* area. The original shapefile was provided by the site related to the data of Lyon Metropolitan Area's actors [data.grandlyon.com]. Since the huge quantity of points contained in the original shapefile (around 190000) would have been very heavy to manage into the GIS, a random extraction of 3000 points have been carried out. The PA population can be seen in figure 2.16. The shape of the points' distribution is due to the administrative boundaries of the *Métropole de Lyon* area.

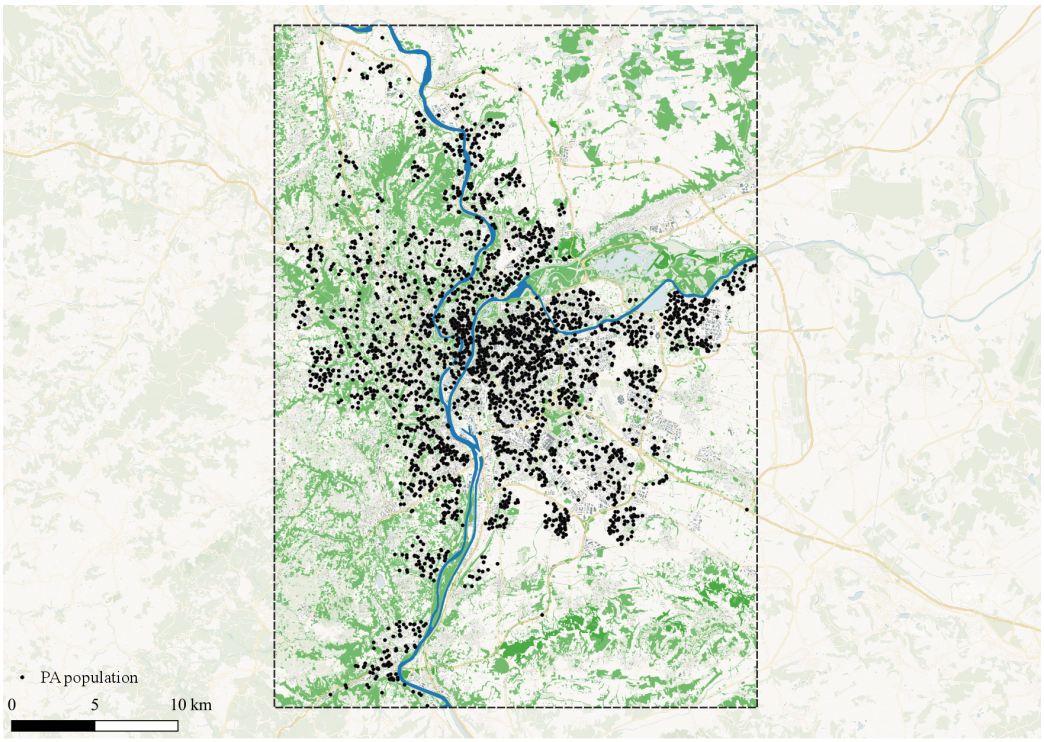

Figure 2.16: Georeference of the PA population subjects into the study domain

2.5.4 Other virtual populations

While the E3N and the PA population are using real addresses, the other three populations created have been only considered in order to check the results. The first two are "semi-random": the amount of points contained in every municipality or neighborhood was selected as a function of the population resident in that area. To have a better distinction into the different areas within the municipality of Lyon (which of course is the biggest and most populated), its area has been further divided, basing on the boundaries of the *conseils de quartier*, in 36 different neighborhoods. The tool that have been used on QGIS was the "Random point creation into polygons", given the polygon of different municipalities into the domain. The last population was fully randomly created within the study domain (figure 2.18b). In order to have a sufficient statistical power, all

those populations counts 2000 subjects.

Figure 2.17: Zoom of the PA population within the downtown

As expected, the fully random population shows some addresses in non-logical places (figure 2.17 vs figure 2.18b). The resolution of the addresses contained in the shapefile of the *Métropole de Lyon* is 0.5 meters (higher than those of all models implemented). The other virtual populations assess the constancy of results obtained with the two set of addresses, and their data will not be showed except if exhibiting an inverse tendency in results.

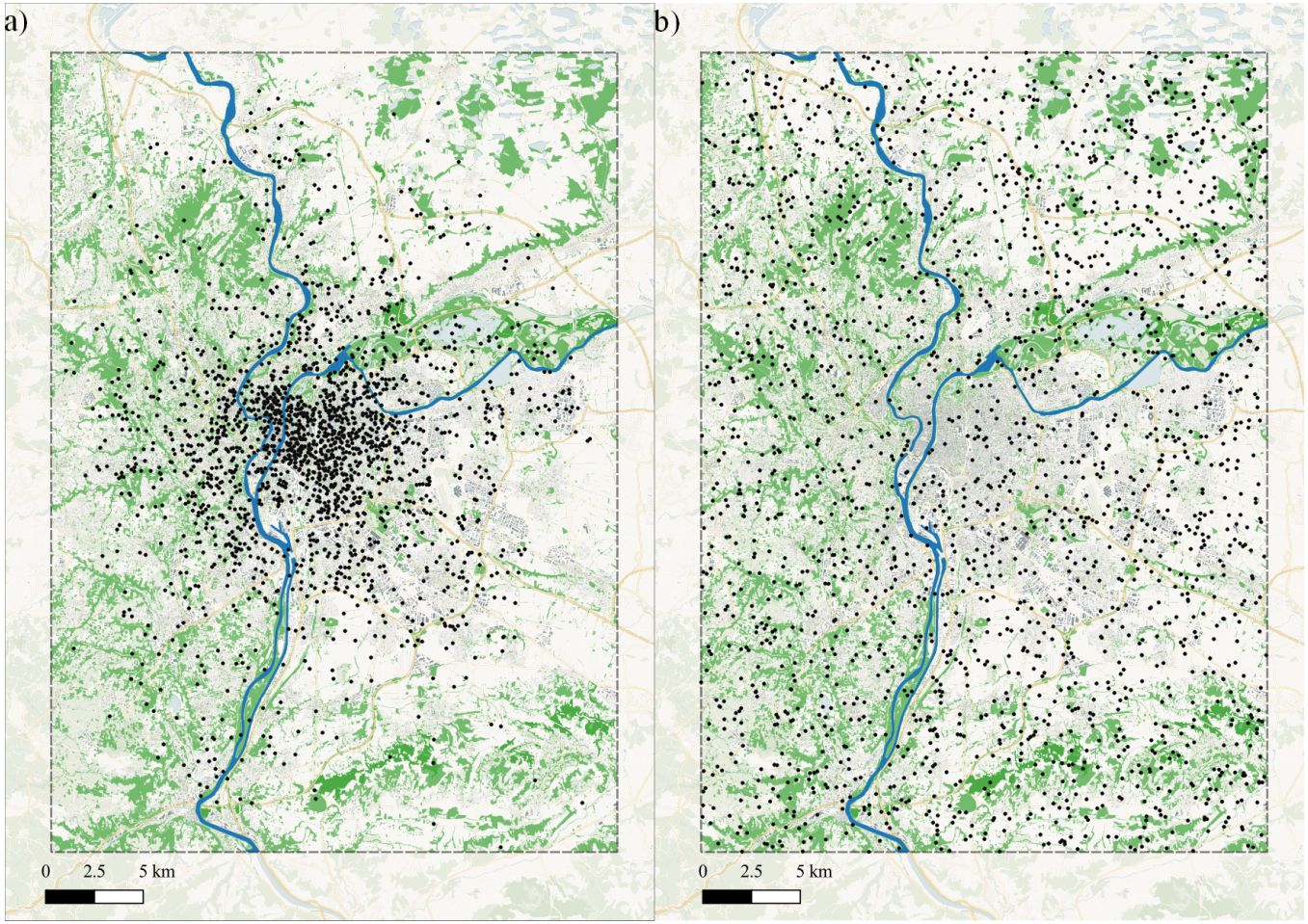

Figure 2.18: Semi-random population (a) and fully random population (b)

2.6 Spatial interpolation and proximity models

2.6.1 Nearest-AQMS and Nearest-CHIMERE modeling

The others modeling approach presented are spatial interpolation models, assigning to each subject an exposure value in function of the distance to a certain point where the pollutant concentration is known or estimated, for example through measurement stations or regional models results. Measurements from the monitoring stations were given by ATMO - Auvergne-Rhône-Alps, the observatory for the air quality surveillance and information of the region recognized by the French Ministry for the ecologic and inclusive transition [https://www.atmo-auvergnerhonealpes.fr]. The Air Quality Monitoring Station network in the city of Lyon is showed in figure 2.19.

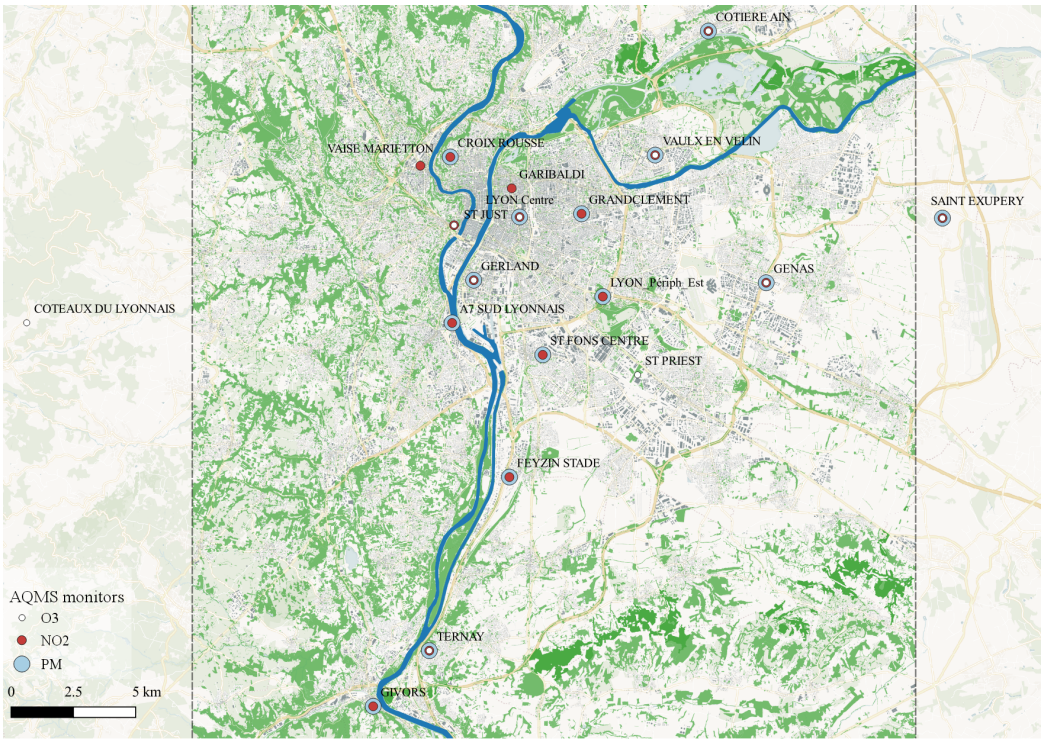

Figure 2.19: AQMS network for the Metropolitan Area of Lyon (ATMO - Auvergne-Rhône-Alps)

In this study, each population's subject is assigned the value of the nearest air quality monitoring station. This has been performed in QGIS 3.4 applying the NNJoin plugin. It is to be noticed that not all AQMS provides measures both for NO_2, O_3 and PM_{10}, consequently it is possible that for the same subject the exposure to two different pollutant is provided by two different AQMS.

The same procedure has been applied to assign exposure values for the CHIMERE model. Actually, CHIMERE's output is inserted into the GIS as a punctual layer, where points are disposed as a grid (figure 2.5). Each point includes the average value for 2010 for NO_2, O_3 and PM_{10}. One of the main interests of this work is to compare data from regional-scale CTM model and AQMS, which are extremely susceptible to small scale-variations, as the number of points on the same domain for the two is similar (20 for CHIMERE, 18 for AQMS) and the exposure-assigning method is the same (assign the value of the nearest point).

2.6.2 Proximity models

GIS are being more and more used in environmental epidemiological studies as a method of exposure assessment based on the residential proximity to distinct types of environmental sources, as traffic roads or industrial facilities. These methods consider the same conceptual approach of a LUR model, further simplifying the exposure assignment procedure by only providing a ranking of the subjects. As exposure to air pollution can be mostly determined by traffic emissions [Colvile et al., 2001], there is a growing evidence that proximity to major roads could be used as a proxy for the exposure to traffic-related air pollution [Miyake et al., 2002] [Venn et al., 2005]. These methods refer both on simple distance-to-road criterion but also on metrics that evaluates the road length in a certain dimension buffer, created around the coordinates of the subjects and intersected with the road network [Hochadel et al., 2006].

The proximity models considered in this study were the distance to nearest road (NEAR), the distance to nearest major road (NEARMAIN)

and the total road length in a 150 meters buffer (BUF150). Data about the road network were provided by the IGN (*Institut National de l'Information Géograohique et forestale*). In the road's vector, each line has an attribute named "importance", classifying the road from 1 to 5, basing on their relative notoriety (French criterion, further information at IGN [2019], pages 319-320). Roads with importance from 1 to 3 were defined as major roads.

Those metrics does not estimate atmospheric concentration and could be compared with the other models only in term of rank correlation.

2.7 Statistical tools for the comparison between the exposure data

The exposure values were estimated by four modelisation approach for the year 2010 (SIRANE, LUR, Nearest-AQMS, Nearest-CHIMERE) and by two (SIRANE and LUR) for the year 2000. Histograms, boxplots and scatterplots with linear regression lines were provided to support the data description and interpretation. The correlation between the estimated values was evaluated through the most widespread statistical indicators applied in literature:

The Pearson correlation coefficient (Pearson's r) is a measure of the linear correlation between two variables. It has a value between -1 and +1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. Mathematically, r value between two variables X and Y is calculated as:

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where $cov(X,Y)$ is the covariance and σ_X and σ_Y the variances of the two distributions.

Spearman's rank correlation coefficient (Spearman's ρ) is a nonparametric measure of rank correlation, i.e. how well the relationship between two variables can be described using a monotonic function. Since the data are converted to ranks, the correlation coefficient does not depend on the actual values and, furthermore, the ranks do not vary if one makes a monotonic transformation of the variables. It can also be defined as the Pearson correlation coefficient between the rank variables, since the calculation formula is the same, only considering X_{rank} and Y_{rank}. Consequently, it also has a value between 1 and -1, where 1 means that the two distributions are ranked exactly in the same way and -1 that they are ranked oppositely.

Cohen's kappa (κ) is a measure of the agreement between a pair of in-

dependent variables evaluating the same phenomena by assigning them ratings. The calculation is based on the difference between how much agreement is actually present (*observed* agreement) compared to how much agreement would be expected to be present by chance alone (*expected* agreement): kappa measures this difference standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate potential systematic disagreement between the observers.

Considering a typical data layout:

		Observer 1		
		Yes	*No*	*Total*
Observer 2	*Yes*	a	b	m_1
	No	c	d	m_0
	Total	n_1	n_0	n

The observed agreement, p_o, is equal to $(a+d)/(n)$, while the expected one is:

$$p_e = \left[\left(\frac{n_1}{n}\right) \cdot \left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right) \cdot \left(\frac{m_0}{n}\right)\right]$$

The value of the Cohen's Kappa (κ) statistic is then calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

In this study, the κ statistic will be applied to evaluate the ability of the model in classifying subject in exposure groups, based on distribution quintiles, and so to assess the models' capability to place the same subject in the same exposure quintile. This indicator can be also defined a measure of the inter-quintile agreement.

The kappa statistic approach can be extended for observers rating more than two categories by means of the disagreement level observed $D_o = 1 - p_o$ and the disagreement level expected $D_e = 1 - p_e$. Let i be the possible ratings for the model A, j the possible ratings for the model B, n_i

the number of subject rated i by A and n_j the number of subject rated j by B. Considering n_{ij} the number of subjects for which the models disagree with A assigning i and B assigning j, the disagreement level observed and expected will be:

$$D_o = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} n_{ij} \qquad D_e = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{n} n_i \cdot n_j$$

And the κ formula:

$$\kappa = \frac{D_e - D_o}{D_e}$$

Finally, the inter-quintile agreement was assessed with interest in weighting differences in ratings assignments proportionally with their dimension. For example, an observation that results in "quintile 1"for model A and "quintile 5"for model B will lower the κ value more than if they were 1-2 or 4-5. This can be calculated by using the weighted kappa statistic ($w\kappa$), which assigns less weight to agreement as categories are further apart. Establishing as v_{ij} the weight for a ij disagreement, the weighted D_o and D_e will be

$$wD_o = \frac{1}{n}\cdot\sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij}\cdot v_{ij} \qquad wD_e = \frac{1}{n^2}\cdot\sum_{i=1}^{n}\sum_{j=1}^{n} n_i\cdot n_j\cdot v_{ij}$$

The $w\kappa$ is then calculated like the base κ.

In this study, Pearson's r, Spearman's ρ and Cohen's $w\kappa$ were calculated for all models' pairs in 2010 and for LUR-SIRANE in 2000.

2.8 Geographical analysis

Among all the different modeling approaches considered in this study, SIRANE and the land use regression model have the finest resolution (10 meters for SIRANE, 50 meters for the LUR) and are the more interesting approaches to consider for further analysis. The objective was to investigate common characteristics between point on which the two models show great differences in the estimation values and investigate the causes, assessing if those differences can depend on the land use type. Furthermore, the two models have been applied to a basic investigation on the relationship between social deprivation factors and exposure levels, which is one of the more widespread typologies of studies that use exposure data into the GIS. The aim of this latter analysis was to search for evident differences in the results provided by the two models.

Land use type analysis

The CORINE Land Cover (CLC) is a standardized data collection on land use type in Europe, aimed to support policy development and being the primary spatial data source on land for European Environmental Agency. CLC is widely used for indicator development, environmental modelling and land cover and land use change analysis in the European context [Büttner, 2014]. The standard CLC nomenclature is hierarchical, including three levels of thematic detail in five major groups (table in Appendix D). The CLC shapefile was applied over the study domain, using the Geographical Information System. Geometries with the same CLC code were merged and the average and median value for the concentration of each pollutant within areas labeled with the same code have been calculated, using the tool "Zonal raster statistics"in QGIS 3.4. The procedure followed was hierarchical: firstly, differences in air pollution average values for codes at level 1 were calculated and displayed in barplots. Then, the same procedure was applied to level 2 and 3 with the aim to narrow it down and highlight ever more specific differences.

Socio-economic factors

Social inequalities related to air pollutant exposure have been widely documented. In these studies, the higher deprivation indices and lower economic positions are usually linked with higher levels of pollutant such as particulate matter and nitrogen oxides [Fairburn et al., 2019]. The inequalities are particularly highly impacting among children: it has been assessed that children living in adverse socio-economic circumstances suffer more often from multiple and cumulative environmental exposures and are likely more susceptible to a variety of toxicants [Bolte et al., 2010]. Morelli et al. [2019] created a "social deprivation heterogeneity coefficient"to be evaluated for ten exposure reduction scenarios for $PM_{2.5}$ in Lyon and Grenoble, defining scenarios on mortality reduction targets and WHO guidelines. They also suggest the example of the Tokyo metropolitan area to demonstrate that strong improvements in air quality likely to entail a large public health benefit can be achieved in large urban areas without compromising mobility.

Most of times, these studies are interested in having a representation of the population via census blocks or neighborhoods of which the average index of socioeconomic factors (average income, social deprivation index, percentage of graduates) are calculated and available in form of vector layer. For this investigation, the deprivation status characteristics were considered at the IRIS level: IRIS represent homogeneous neighborhoods and are the finest geographical census unit available in France. They are similar to the US census block group and contain on average 2000 inhabitants, even if some IRIS located outside of the town can count less people. Within this study domain, 609 IRIS are present.

For each IRIS, the mean value of PM_{10} and NO_2 was calculated, both for SIRANE and LUR, for the year 2010 through the tool "Zonal Raster statistics"in QGIS3.4. IRIS data were ranked according to their average income value and then were split into five equally sized groups (quintiles). Boxplots showing the variation of the sampled distribution were plotted and discussed both for SIRANE and LUR, focusing on similarities and dif-

ferences between the two models' results. Also, a statistical comparison between the groups (Wilcoxon test) have been applied for both models with interest in finding if the same conclusions can be taken for SIRANE and LUR about the inter-group variability. Furthermore, the Wilcoxon test has been applied between residuals SIRANE-LUR distributions of the different quintiles, to investigate if the differences between the two models are systematically dependent on the average incomes.

2.9 Odds ratio comparison

The main objective of environmental epidemiological studies is to quantify if there is a significant risk of an adverse effect associated with a specific exposure. In the framework of the XENAIR project, that effect is an increased risk of breast cancer, for women most exposed to air pollution. The reference indicator is the *Odds Ratio*. Considering a typical study layout, it is defined as follows:

	Diseased	Healthy	Total
Exposed	a	b	e_1
Not exposed	c	d	e_0
Total	m_1	m_0	N

$$OR = \frac{a \cdot d}{c \cdot b}$$

That corresponds to the ratio between the absolute risks for the exposed ones (a/b) and for the not exposed ones (c/d).

The setup for the XENAIR project studies is the nested case-controls one, implicating that cases (subjects manifesting adverse health effects) and controls (sane subjects) are present in 1:1 ratio. Being exposed is associated with the disease occurrence in presence of a OR >1 with statistical significance at a certain confidence level (most of times 95%). The statistical significance is verified by chi-squared independence test, with OR=1 as null hypothesis and OR >1 as alternative hypothesis. The statistic value is equal to:

$$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{e_1 \cdot e_0 \cdot m_1 \cdot m_0}$$

To accept OR >1, the χ^2 must be over the correspondent value on the chi-squared distribution with 1 degree of freedom and p=0.95, that is equal to 3.841. In order to provide a simpler visualization of the statistical significance of the results, odds ratios are usually showed with their corre-

spondent confidence interval, calculated according to the Woolf method [Woolf, 1955]:

$$CI = \left[OR \cdot \exp\left(\pm z \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right) \right]$$

Where z is the standard normal distribution value corresponding to the confidence value needed (1.96 for 0.95). Consequently, the statistical significance at a 95% confidence level is attributable to odds ratios whose confidence interval is set completely above the unity, that corresponds to a χ^2 value above 3.841.

Usually, epidemiological studies about air pollution effects on human health are based on the categorization of exposure values in groups ("less exposed", "highly exposed"), so that the odds ratios are calculated regarding inter-groups variability. These categorizations are often defined basing on the exposure values distribution quantiles: for example, Danjou et al. [2019] evaluated the OR associated with exposure quintiles of dioxins and breast cancer occurrence. Other studies perform a categorization into quartiles [Andersen et al., 2017b], [Gray et al., 2010], [Nie et al., 2007].

Assuming SIRANE results as a reference, the performances of LUR in estimating inter-quartile odds ratio (i.e. between four exposure groups defined by the distribution quartiles) have been evaluated through an iterative procedure for the year 2010. A virtual case-control cohort counting 10000 subjects within the study domain was built, assigning cases and controls so that their distribution between the exposure groups estimated by SIRANE lead to a given result in term of odds ratio, set as equal to a reference. The number of subjects is the same of the E3N cohort involved into the XENAIR project. Then, replacing SIRANE's values with those given by the LUR model to the same case-control cohort, the odds ratios were calculated considering LUR's exposure groups and compared with the SIRANE ones.

The aim is to quantify the loss of statistical significance due to the use of the LUR model instead of SIRANE. Table 2.3 shows the reference inter-

quartile subjects' repartition and ORs, set basing on typical values for air pollution - breast cancer association provided by epidemiologists working at the *Leon Bérard* center. Q1 stand for "Quartile 1"and identifies the less exposed group, in the way that the odds ratios for the others are calculated with respect to it. For that reason, the OR for Q1 is always equal to 1.

	Cases	Controls	Total	OR	CI95%
Q4	1364	1250	2614	1.20	1.07 - 1.34
Q3	1280	1250	2530	1.13	1.01 - 1.26
Q2	1220	1250	2470	1.07	0.96 - 1.20
Q1	1136	1250	2386	1.00	-
Total	5000	5000	10000		

Table 2.3: Cases and controls reference repartition into the cohort

This original procedure has been developed since a consolidated method for such a comparison has not been found in the literature. It develops through the following steps and a graphical explanation is given in Appendix E.

- The totality of SIRANE cells within the domain was sampled, resulting in a table containing more than 12 million rows. Each row was associated to a SIRANE exposure value, a LUR exposure value and to a value referring to the population density;
- Population density-weighted quartiles were calculated, both for SIRANE and LUR distributions, and all the table rows were categorized in four exposure groups for the two models depending on quartiles' values. Each table row is considered as a potential virtual subject, assigned to an exposure group for SIRANE (from 1 to 4) and for LUR (from 1 to 4);
- 500 random sampling of 10000 virtual subjects were made by imposing the case-control inter-quartile distribution presented in table 2.3 with respect to the SIRANE values. Each simulation step results

in a population of 10000 subjects whose case-control distribution between the four exposure groups is always equal to the reference for SIRANE while it is different at each step for LUR;

- For each sampling step, the odds ratios for Q2, Q3 and Q4 were calculated with respect to the LUR-defined groups for the same case-control cohort, resulting for both Q2, Q3 and Q4 in 500 odds ratios estimated with their confidence interval (95%).

All the procedure was archived by coding a specific function in RStudio, mainly involving tools from the packages "dplyr"and "epitools". The assessment was made both for NO_2 and PM_{10} for the year 2010.

Pollutant concentration quartiles were defined as weighted by the population density to mostly focus on populated areas, grouping into the same quartile (the first) a wide range of low exposure values set in places where almost nobody lives and consequently of low epidemiological interest. Data about the population density were provided by the *Institut National de la Statistique et des Études Économiques* (INSEE) [https://www.insee.fr/fr/statistiques /4176290].

Average odds ratio resulting from the simulations for the LUR model were compared with the reference. The percentage of the simulations that maintained a significant value for Q3 and Q4 was calculated and discussed. In addition, the same procedure was applied to a subset of highly exposed subjects, applying the whole procedure considering only cells with values over a defined threshold for SIRANE (25 $\mu g/m^3$ for NO_2 and 20 $\mu g/m^3$ for PM_{10}), chosen in order to limit the domain to the boundaries of the principal urban area.

3 Results and discussion

3.1 Year 2010

3.1.1 Results for the real population

The data analysis is presented both for the real population (E3N cohort) and for the *Points d'addressage* population. Table 3.1 ad figure 3.1 present the summary statistics for the three pollutants (NO_2, O_3, PM_{10}) and four models (SIRANE, LUR, Nearest-AQMS, Nearest-CHIMERE).

For the NO_2, SIRANE and LUR have both similar central values and

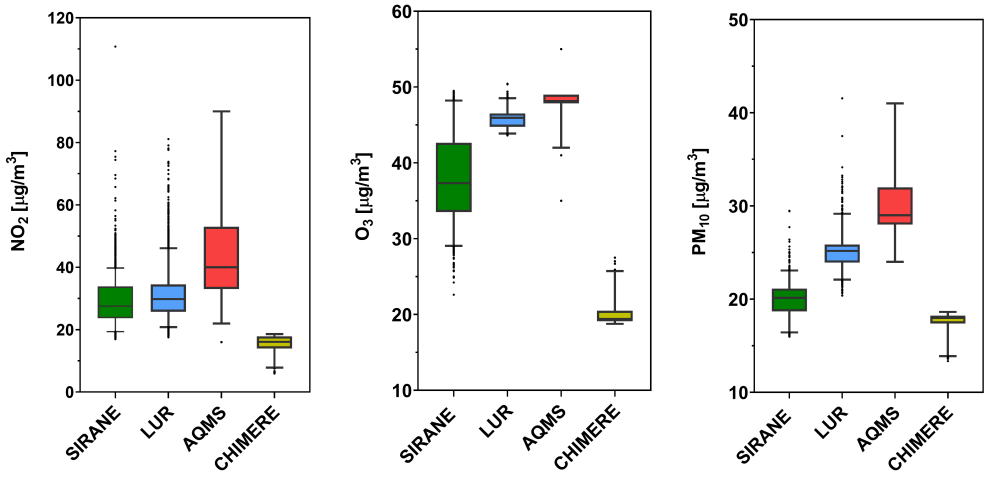

Figure 3.1: Boxplots for the real population exposures in 2010

	SIRANE	LUR	Nearest AQMS	CHIMERE
			NO_2 [$\mu g/m^3$]	
Min	16.10	15.96	16.00	5.79
1^{st} quartile	25.21	26.26	34.00	15.56
mean	31.85	33.19	47.67	15.98
median	32.69	31.96	40.00	17.85
3^{rd} quartile	38.13	37.37	62.00	17.85
95^{th} quantile	45.67	50.82	90.00	18.57
Max	62.12	105.97	90.00	18.57
SD	8.71	9.93	18.80	3.37
			O_3 [$\mu g/m^3$]	
Min	22.62	43.62	35.00	18.77
1^{st} quartile	33.52	44.76	48.00	19.26
mean	38.14	45.85	47.73	20.39
median	37.32	45.94	48.00	19.26
3^{rd} quartile	42.63	46.52	49.00	20.50
95^{th} quantile	48.19	48.52	49.00	25.73
Max	49.50	50.44	89.00	27.51
SD	5.85	1.34	6.37	2.14
			PM_{10} [$\mu g/m^3$]	
Min	15.97	20.38	24.00	13.36
1^{st} quartile	18.69	23.93	28.00	17.39
mean	19.97	25.28	29.98	17.63
median	20.13	25.17	29.00	18.11
3^{rd} quartile	21.14	25.83	32.00	18.11
95^{th} quantile	23.07	29.13	41.00	18.64
Max	29.49	41.55	41.00	18.64
SD	1.96	2.29	4.53	1.26

Table 3.1: Data description for the real population exposures in 2010

variability (sd equals to 8.71 and 9.93). The minimums and the first quartiles are almost the same (around 16 $\mu g/m^3$ and 25.5 $\mu g/m^3$, respectively), while a difference is observed for the highest values, where the LUR has more outliers and a higher 95th quantile (50.82 vs 45.67 $\mu g/m^3$). The Nearest-AQMS presents data with the highest variability (sd=18.80), probably because those values are strongly influenced by conditions in the close proximity of the monitoring station.

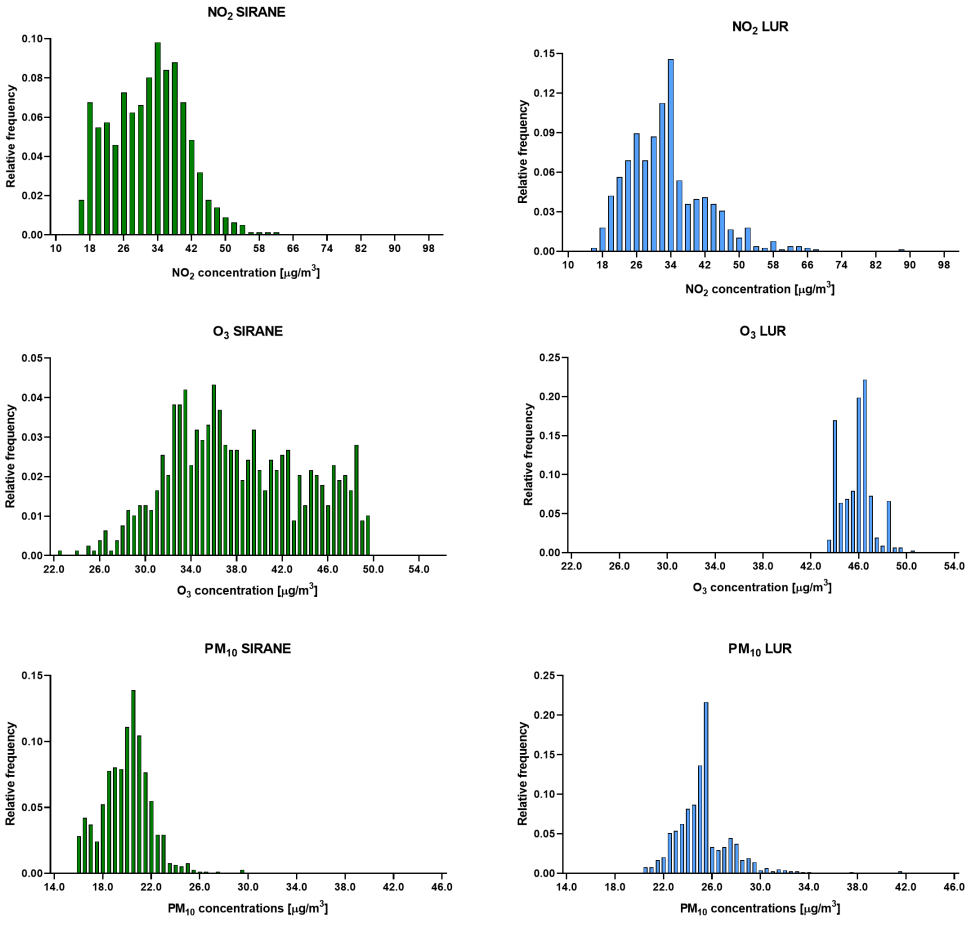

Figure 3.2: Histograms representation for the real population, SIRANE and LUR

NO_2 distributions are, both for SIRANE and LUR, heavy-tailed on the right, with most of values concentrated in the left part of the graph (range

between 20 and 50 $\mu g/m^3$) while the higher values are quite less and distant from each other. This was already observed for NO_2 spatial distribution in urban environment [Beelen et al., 2010], which is characterized by peaks of concentrations in correspondence of main roads. Considering the NO_2 limit for the annual average exposure provided by the WHO guidelines, which is 40 $\mu g/m^3$, the percentage of subjects exposed over this threshold is comparable between the two models (16.69 % for SIRANE, 19.75 % for LUR) [WHO, 2005].
For the PM_{10}, different models are all set into different ranges. Comparing SIRANE and LUR, data have about the same variability: inter-quartile ranges are both slightly over 1 $\mu g/m^3$ but placed in different ranges (18.69-20.13 $\mu g/m^3$ for SIRANE, 23.95-25.17 $\mu g/m^3$ for LUR). Boxplots (Figure 3.1) clearly shows that LUR values are quite higher than SIRANE's. The fact that the result of SIRANE EXT for PM_{10} has a background concentration value far inferior respect to the measured one (see section 2.3) certainly affects this characteristic. The percentage of exposure values above the WHO guidelines limit for PM_{10} annual average exposure (20 $\mu g/m^3$) is 53.38 % for SIRANE, while for the LUR the totality of subjects experience an air quality level exceeding the limit [WHO, 2005]. Boxplots shows a weak agreement between models for O_3. Except for SIRANE, all IQRs are quite low and so all boxplots are squeezed into very small ranges. The standard deviation of LUR is very low (1.34 $\mu g/m^3$), suggesting a quite poor spatial variation among the domain (all values are included between 43.52 and 50.44 $\mu g/m^3$, while for SIRANE the range is 22.62 - 49.50). The Nearest-AQMS model has a very small resolution with only 6 ozone stations all over the city, and consequently all the expositions are limited over those 6 values (This explains why 75th and the 95th quantile are both equal to 49 $\mu g/m^3$).

Figure 3.3 shows paired scatterplot for the different model exposure estimations. It is clearly observable that the LUR and SIRANE models are the ones that have the best intercorrelation, while the spatial interpolation models do not have a sufficient resolution to guarantee a precision in pollutant estimation at a few-meters scale.

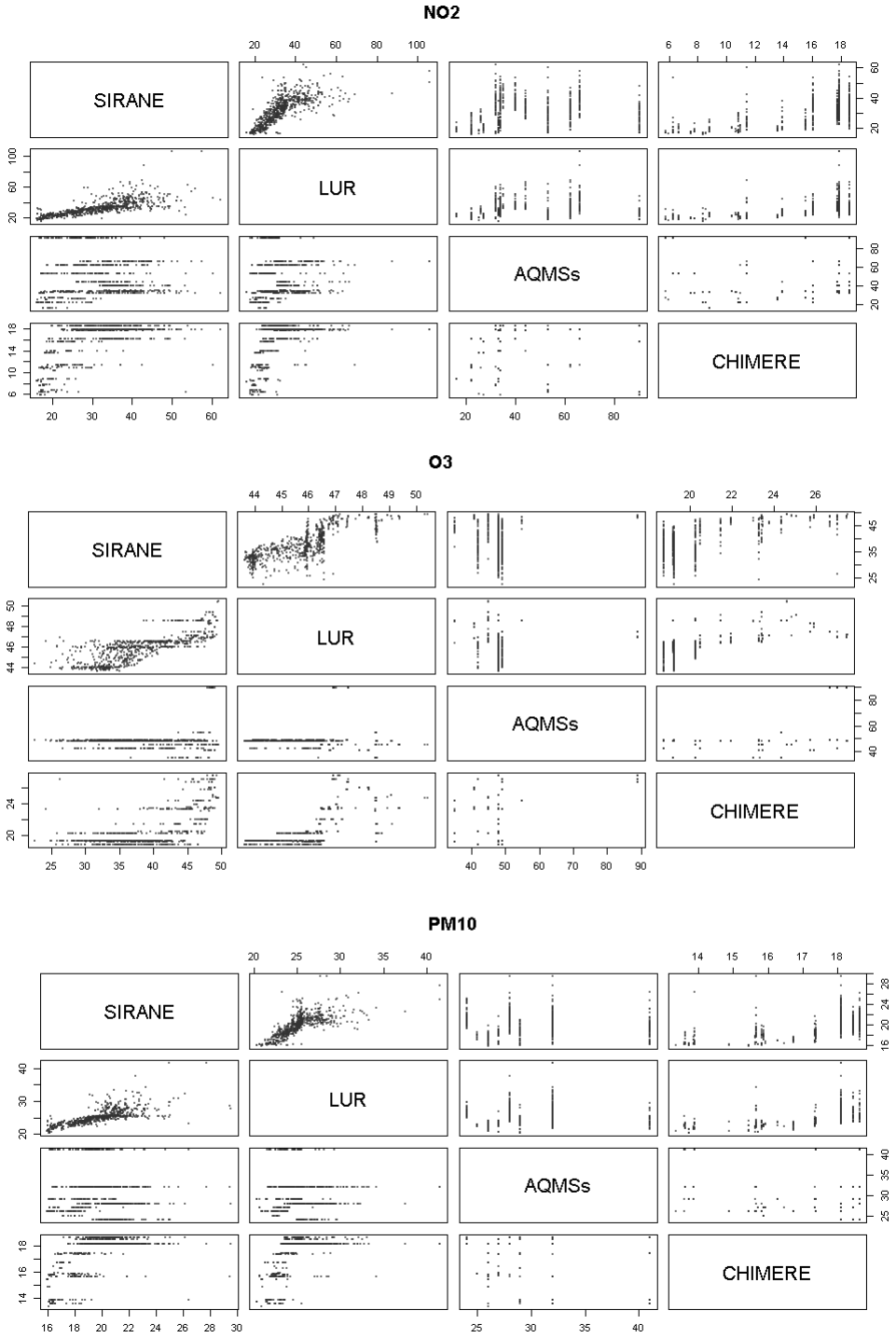

Figure 3.3: Paired-wise scatterplots of air pollution exposition estimated for the real population from the four models in 2010

Since the exposure assignment method for Nearest-AQMSs and CHIMERE results in having an amount of values (measurement stations or CHIMERE points) that is quite lower than the number of subjects, lots of points are assigned to the same value (see figure 3.3). Similar patterns were observed by Marshall et al. [2008], who investigated the correlation between a Nearest-AQMS station, a LUR model and a dispersion model in Vancouver. On the other hand, LUR and SIRANE result in having a visual positive correlation. Linear regression has been performed over those two models, selecting LUR as independent variable and SIRANE as dependent, for the three pollutants (figure 3.4), whose coefficients are resumed in table 3.2.

	Intercept	Slope	R^2
NO_2	10.53	0.64	0.54
O_3	-116.49	3.37	0.59
PM_{10}	4.42	0.62	0.51

Table 3.2: Coefficients and adjusted coefficients of determination of linear regression lines for the real population, LUR vs SIRANE

For all pollutants it is observable from R^2 values that the variance SIRANE values can be well predicted from LUR's: assuming SIRANE as a reference model for urban prediction modelling at high resolution, it is interesting that LUR explains at least 50 percent of its variance, being considered (at least for preliminary considerations) a valuable predictor of SIRANE outputs.

Figure 3.4 shows that LUR values for PM_{10} are located in a clearly higher range than SIRANE's. It is evident that only a few points are above the bisector, so only for those subjects (3 in 785) the exposition to PM_{10} is higher for SIRANE than for LUR. The ozone scatterplot exhibits a quite dissimilar frame, due to the fact that LUR presents a very limited data spectrum. Because of this, the linear regression line is more inclined than the others (slope=3.30).

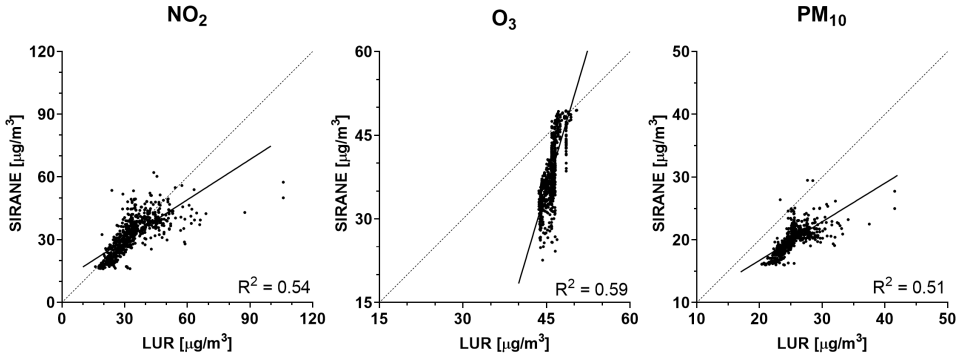

Figure 3.4: Linear regression lines for the real population between LUR and SIRANE; the dashed line represents the bisector

In figure 3.5, the real population is showed grouped in 5 different NO_2 exposure classes for the four models. For SIRANE and LUR, the most exposed subjects are placed near or within the city center, as expected, and a sufficient accuracy in differently classifying close points can be noted. The Nearest-AQMS model shows a great tendency to misclassify subjects who do not have a monitor in their immediate proximity and are assigned to far ones, for example subjects addressed on the hills at the western part of the domain. The model CHIMERE confirms to be unable to estimate precise concentration values and, with respect to other models, clearly underestimates the exposures.

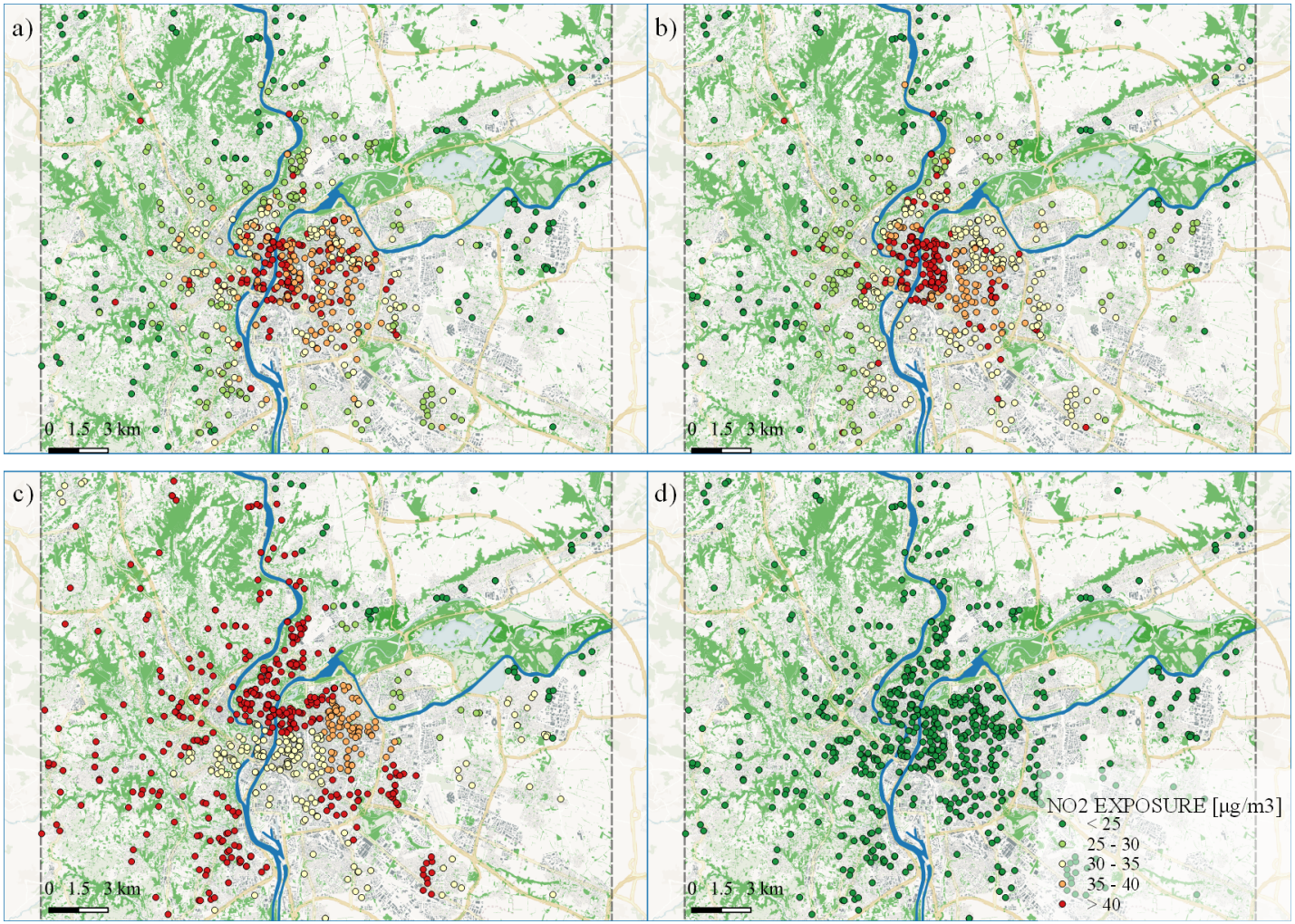

Figure 3.5: Exposure classification of the real population in Lyon; a) SIRANE ; b) LUR; c) Nearest-AQMS; d) Nearest CHIMERE

E

3.1.2 Results for the PA population: differences and similarities

All analyses have been also conducted for the PA population. Even if the results are very similar, some differences have been observed (table 3.3).

For the NO_2, the biggest differences emerging by a data analysis are the maximum values for the SIRANE model, which in the PA population are quite higher than in the real. This is related to a characteristic of the shapefile at source of this population, that comprises not only civil building addresses but also a few points along main roads, situated on highways or at urban junctions. Actually, 10 of such points are present (0,33% of the total PA population), including the most exposed one (NO_2 that is equal to 110.79 $\mu g/m^3$ and situated on a highway in the common of *La Mulatière*, south of Lyon). Skipping those points, the maximum value is 69.6 $\mu g/m^3$, which is a little higher than the real population's maximum (62.12 $\mu g/m^3$) due to the fact that the PA's subjects, counting many more subjects (3000 vs 785), of course have a higher probability to sample very high concentrations.

The other major difference that can be seen is related to the maximum value for ozone into the Nearest-AQMS model, which for the PA population is quite less (55 instead of 89 $\mu g/m^3$). This is due to the shape of the *Métropole of Lyon* area: as it can be clearly seen by figures 2.16 and 2.19, the AQMS of *Coteaux du Lyonnais*, which is the one giving the highest O_3 concentration, is very far from the boundary of the metropolitan area. For the PA population, no subject was associated with that AQMS, while 14 subjects of the real population were.

The O_3 histograms reported in figure 3.6 for SIRANE show a clear tendency to a bimodal shape, probably due to the presence of two different clusters of data relative to rural and urban concentration values. This tendency is slightly observable also in the LUR histogram. Other relevant differences are not present, if not the ones simply due to greater numeros-

	SIRANE	LUR	Nearest AQMS	CHIMERE
			NO_2 [$\mu g/m^3$]	
Min	16.95	17.55	16.00	5.97
1^{st} quartile	23.66	25.70	33.00	13.92
mean	28.84	31.18	44.44	15.50
median	27.56	29.82	40.00	16.08
3^{rd} quartile	33.87	34.50	53.00	17.85
95^{th} quantile	39.76	46.05	90.00	18.57
Max	110.79	81.11	90.00	18.57
SD	7.09	8.27	18.62	3.45
			O_3 [$\mu g/m^3$]	
Min	13.14	43.57	35.00	18.77
1^{st} quartile	36.30	45.90	45.00	19.21
mean	40.14	46.28	46.52	20.67
median	40.89	46.40	48.00	20.27
3^{rd} quartile	43.76	46.58	48.00	21.46
95^{th} quantile	46.89	48.53	49.00	25.73
Max	48.83	50.34	55.00	27.51
SD	4.70	1.20	3.92	2.19
			PM_{10} [$\mu g/m^3$]	
Min	16.18	20.62	24.00	13.36
1^{st} quartile	18.41	23.57	27.00	17.36
mean	19.43	24.83	29.89	17.50
median	19.17	24.73	29.00	18.11
3^{rd} quartile	20.36	25.50	32.00	18.45
95^{th} quantile	21.61	28.09	41.00	18.64
Max	45.26	36.20	41.00	18.64
SD	1.63	1.92	4.39	1.34

Table 3.3: Data description for the PA population in 2010

ity of the sample. The main consequence of this latter aspect is the shape of the histograms, that further tends to a heavy-tailed distribution for NO_2

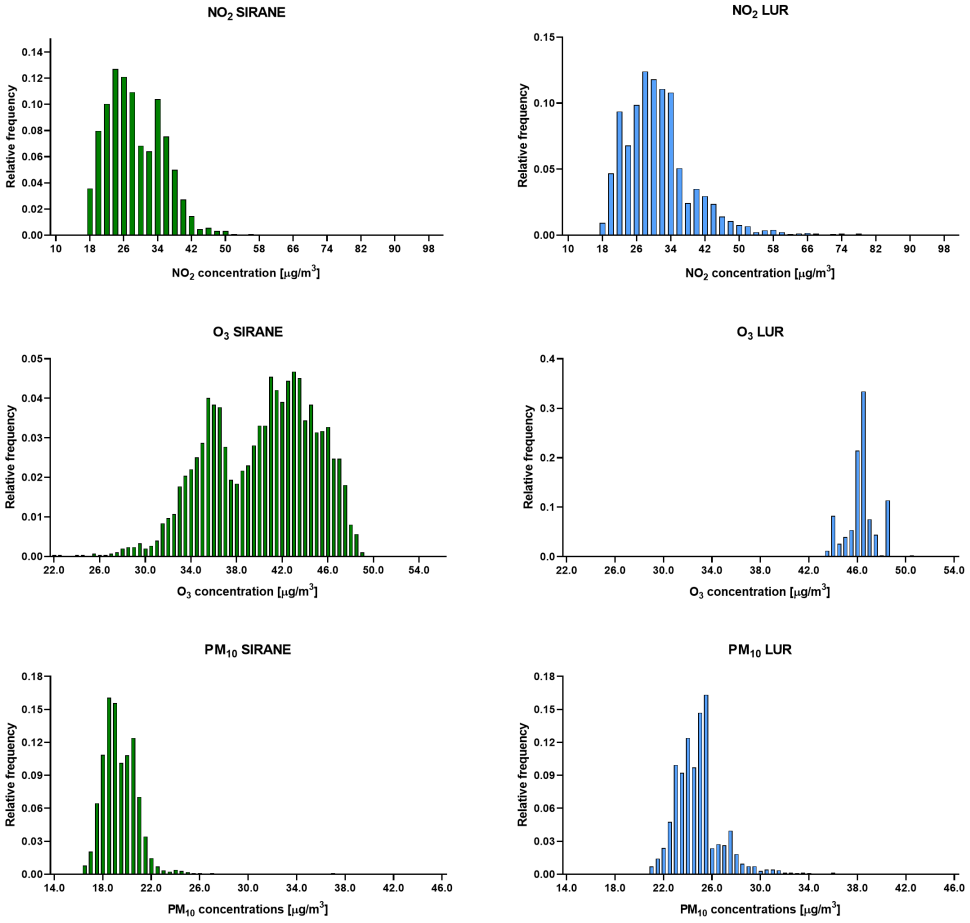

Figure 3.6: Histograms representation for the PA population, SIRANE and LUR. Since this population counts more subjects, the right-tailed shape of the distribution for NO_2 and PM_{10} is accentuated with respect to the real's one, figure 3.2.

and PM_{10} in SIRANE and LUR models.

3.1.3 Correlations and agreement coefficients

Data have been further processed in order to assess their correlations and their agreement level. Pearson's r, Spearman's ρ and Cohen's $w\kappa$ have been calculated for the three pollutants and four modeling approaches. It

is appropriate to recall that Pearson's r evaluates the linear correlation between the concentration values estimated by two models, whereas Spearman's ρ and Cohen's $w\kappa$ assess their rank correlation (regarding Spearman's ρ) and their inter-rater reliability (regarding $w\kappa$).

Table 3.4 shows the Pearson's coefficient for the real population and for all couples of models.

	NO_2	O_3	PM_{10}
SIRANE - LUR	0.73	0.77	0.72
SIRANE - Nearest AQMS	0.04	0.09	-0.11
SIRANE - CHIMERE	0.62	0.64	0.58
LUR - Nearest AQMS	0.07	-0.04	-0.04
LUR - CHIMERE	0.51	0.60	0.51
CHIMERE - Nearest AQMS	0.18	0.31	-0.05

Table 3.4: Pearson's r for the real population, 2010

It is observed that the correlation level between two models does not vary considerably depending on the pollutant examined. Overall, the stronger correlation is seen between SIRANE and LUR (0.72 - 0.77), but there are acceptable values also for SIRANE-CHIMERE and LUR-CHIMERE. The lowest correlations are therefore the ones that include the Nearest-AQMS model, that only shows reasonable values with CHIMERE (0.31 for the O_3).

High correlation values for NO_2 between a dispersion model and a LUR have been observed in various studies. Cyrys et al. [2005] calculated a very high Pearson coefficient (0.83) in Munich between a LUR and a Gaussian multisource dispersion model. Wang Meng et al. [2015] calculated an even higher correlation (0.9) in studying the relationship between air pollution and lung function in children in the Netherlands, considering a LUR and a dispersion model. The very low values of Pearson's r for SIRANE-Nearest AQMS and LUR-Nearest AQMS are almost certainly caused by the poor density of monitors. Indeed, Sellier et al. [2014] performed a similar analysis with 54 AQMSs in Nancy and Poitiers, over 2002

women members of the EDEN mother-child cohort. They observed quite higher values of r both for Dispersion Modelling - Nearest AQMS (r = 0.63) and for LUR - Nearest AQMS (r = 0.54), even if also in their study the correlation between Dispersion Model - LUR was still greater (over 0.7).

	NO_2	O_3	PM_{10}
SIRANE - LUR	0.84	0.80	0.82
SIRANE - Nearest AQMS	0.15	-0.15	-0.10
SIRANE - CHIMERE	0.54	0.48	0.51
LUR - Nearest AQMS	0.14	-0.16	-0.07
LUR - CHIMERE	0.51	0.51	0.52
CHIMERE - Nearest AQMS	0.33	0.06	0.28

Table 3.5: Spearman's ρ for the real population, 2010

Table 3.5 shows the Spearman's rank correlation coefficients for the real population. A general constancy of the correlation between two models for different pollutants is still observed. Moreover, values of ρ are higher than Pearson's for the SIRANE-LUR comparison, all being over 0.8 and meaning a quite strong statistical association between the rankings of these two distributions. Figure 3.7 shows scatterplots for the NO_2 between SIRANE-LUR, SIRANE-CHIMERE and SIRANE-Nearest AQMS.

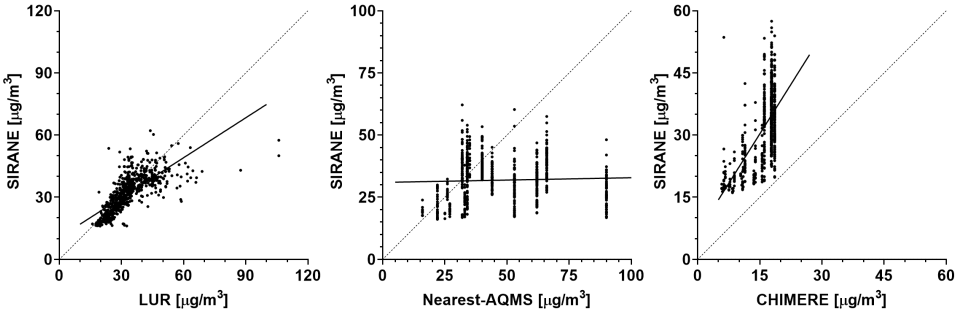

Figure 3.7: NO_2 scatterplot for SIRANE-LUR, SIRANE-CHIMERE, SIRANE-Nearest AQMS; the dashed line is the bisector

A very few points are ranked in different way by SIRANE and LUR, as

it is shown in figure 3.7. This is not true for the CHIMERE and even less for the Nearest-AQMS. Anyway, despite considerably underestimating the exposure values (compared to SIRANE) in a systematic way, CHIMERE shows a weak correlation with SIRANE (ρ = 0.54) that is slightly viewable into the figure 3.7. Good rank-correlations between LUR and dispersion modelling for NO_2 and PM_{10} are quite diffused in literature. For example, in a study on commuter exposure to NO_2 in Basel, Spearman coefficents were observed almost equal to 1.0 in comparing Gaussian-type dispersion models and a LUR european model (ESCAPE) [Ragettli et al., 2014].

Weighted Cohen's Kappas ($w\kappa$) have been calculated to estimate the level of inter-rater reliability between the various models. Exposition values for all models has been categorised in quintiles according to their exposure value: 5 is assigned to the 20% most exposed subjects, 1 to the 20% lowest and so on. Once a value from 1 to 5 was assigned to all subjects (for 4 modeling approaches and 3 pollutants), $w\kappa$ were computed by an appropriate function coded by cycling the "ckap"function in R, from the package "rel". Results are summarized in table 3.6. As $w\kappa$ is an estimated value, it is fundamental to also report a confidence interval: 95% confidence was chosen as the most diffused in literature.

models	NO_2	95% CI	O_3	95% CI	PM_{10}	95% CI
SIRANE - LUR	0.66	0.63 - 0.69	0.61	0.58 - 0.65	0.64	0.6 - 0.67
SIRANE - Nearest AQMS	0.08	0.03 - 0.13	-0.15	-0.19 - -0.1	-0.04	-0.08 - 0
SIRANE - CHIMERE	0.37	0.33 - 0.41	0.18	0.13 - 0.22	0.35	0.31 - 0.39
LUR - NearestAQMS	0.08	0.03 - 0.13	-0.14	-0.18 - -0.1	-0.05	-0.09 - -0.01
LUR - CHIMERE	0.36	0.32 - 0.41	0.14	0.1 - 0.19	0.33	0.29 - 0.38
CHIMERE - NearestAQMS	0.34	0.28 - 0.39	0.26	0.2 - 0.32	0.27	0.23 - 0.31

Table 3.6: Weighted Kappas ($w\kappa$) for the real population

In table 3.6, colors refer to the classification made by Viera and Garrett [2005], that classifies values between 0.1 and 0.2 as "slight agreement"(black), 0.2-0.4 as "fair agreement"(red), 0.4-0.6 as "moderate agreement"(yellow) , 0.6-0.8 as "substantial agreement"(blue) and 0.8-1.0 as "almost perfect agreement"(green).

The best reliability is observed between LUR and SIRANE, as expected. They show a substantial agreement for all pollutants, with $w\kappa$ values between 0.61 and 0.66. Similar values were observed by Coudon et al. [2019a] for a comparison between SIRANE and several GIS-based metrics. On the other hand, lower values are observed for the SIRANE - Nearest AQMS and LUR - Nearest AQMS, confirming the already suggested idea that comparing models with very different resolution leads to poor correlation and agreement. This was already seen in table 3.4 for the Pearsons' r and in table 3.5 for the Spearman's ρ, where values between SIRANE-LUR were the highest and values between SIRANE/LUR and Nearest AQMS were the lowest.

Regarding the CHIMERE model performances, the quite low resolution (20 over the whole domain, as said) is partially balanced by the spatial homogeneity of the points (see figure 2.5). Moreover, the fact of being a regional model ensures that estimated values are not too much affected by local variations (indeed, as seen in chapter 3.1.1, their range is very limited). The overall incapability of estimating precise concentration values is so partially balanced by the capacity of correctly distinguish areas with higher and lower concentrations, so that subjects are ranked in a more appropriate way with respect to the Nearest-AQMS method. The graphs and the correlation coefficients calculated for the virtual populations did not show a different tendency for the results. However, they are presented in Appendix A and B.

3.1.4 Correlations and agreement coefficients for the PA population: differences and similarities

The results of the same analysis on the PA population do not show big differences. Pearson's r (table 3.7) are almost the same seen in table 3.4, with a little decrease in the values for the pairs LUR-CHIMERE and SIRANE-CHIMERE.

	NO_2	O_3	PM_{10}
SIRANE - LUR	0.75	0.71	0.70
SIRANE - Nearest AQMS	0.11	-0.13	-0.01
SIRANE - CHIMERE	0.58	0.61	0.48
LUR - Nearest AQMS	0.06	-0.29	0.02
LUR - CHIMERE	0.53	0.49	0.51
CHIMERE - Nearest AQMS	0.20	0.11	0.00

Table 3.7: Pearson's r for the PA population, 2010

As for the Pearson's r, also for the Spearman's ρ the main difference seen is associated to SIRANE-CHIMERE and LUR-CHIMERE, but with higher values for the PA population (table 3.8 vs table 3.5). These small variations are probably related to a simple scale effect.

	NO_2	O_3	PM_{10}
SIRANE - LUR	0.75	0.71	0.70
SIRANE - Nearest AQMS	0.11	-0.13	-0.01
SIRANE - CHIMERE	0.58	0.61	0.48
LUR - Nearest AQMS	0.06	-0.29	0.02
LUR - CHIMERE	0.53	0.49	0.51
CHIMERE - Nearest AQMS	0.20	0.11	0.00

Table 3.8: Spearman's ρ for the PA population, 2010

Cohen's weighted kappas (table 3.9) also shows little differences due to sample numerosity, above all regarding the 95% confidence intervals,

which are quite smaller than the those of the reals (table 3.6). So, as expected, $w\kappa$ estimated for the PA population are more accurate.

models	NO_2	95%CI	O_3	95%CI	PM_{10}	95%CI
SIRANE - LUR	0.68	0.66 - 0.69	0.59	0.57 - 0.61	0.67	0.65 - 0.68
SIRANE - Nearest AQMS	0.14	0.11 - 0.16	-0.09	-0.11 - -0.07	0.06	0.04 - 0.09
SIRANE - CHIMERE	0.43	0.41 - 0.45	0.35	0.33 - 0.38	0.38	0.36 - 0.40
LUR - NearestAQMS	0.11	0.08 - 0.13	-0.05	-0.07 - -0.03	0.09	0.07 - 0.12
LUR - CHIMERE	0.41	0.39 - 0.43	0.31	0.29 - 0.33	0.38	0.36 - 0.40
CHIMERE - NearestAQMS	0.32	0.29 - 0.34	0.21	0.18 - 0.24	0.20	0.17 - 0.23

Table 3.9: Cohen's Kappas (wκ) for the PA population, 2010

In addition, values of $w\kappa$ are slightly higher both for SIRANE-CHIMERE and LUR-CHIMERE. SIRANE-LUR shows better agreement for NO_2 (0.68 vs 0.66) and PM_{10} (0.67 vs 0.64), while for O_3 it decreases (0.59 vs 0.61).

3.1.5 Proximity models

Correlation and agreement coefficients were evaluated also for the proximity models whose objective is to estimate the exposure level of the subjects basing on their proximity to roads. These methods do not provide values of concentrations but are useful tools during preliminary analysis prior to more sophisticated research [Zou et al., 2009]. As explained in section 2.6.2, three indicators have been chosen: distance to the nearest road (NEAR), distance to the nearest main road (NEARMAIN) and road length sum in a 150 meters buffer (BUF150). The correlation with NEAR and NEARMAIN was evaluated computing the inverse of the distance.

The correlation was assessed in term of ranking capability and inter-rater agreement between the proximity values and SIRANE/LUR. Nearest-AQMS and CHIMERE models were not considered for this analysis. Moreover, only primary traffic pollutant (NO_2 and PM_{10}) were evaluated. Table 3.10 shows Spearman-rank correlation between LUR/SIRANE and the proximity models.

	NO_2	PM_{10}		NO_2	PM_{10}
SIRANE - NEAR	-0,11	-0,10	LUR - NEAR	-0,17	-0,17
SIRANE - NEARMAIN	0,44	0,44	LUR - NEARMAIN	0,49	0,48
SIRANE - BUF150	0,43	0,39	LUR - BUF150	0,50	0,47

Table 3.10: Spearman's ρ between SIRANE/LUR and proximity models

Values calculated for Spearman's ρ shows the absence of correlation between both pollutants and the NEAR metrics, indicating the proximity to a road as an unreliable indicator. The result was expected and justified by the fact that, including into the analysis all type of roads, the proximity to small streets with little traffic or even partially pedestrian paths is clearly leading to the misclassification of most of the subjects. On the contrary, the proximity to a major road were found as a stronger model (0.44 for SIRANE both in NO_2 and PM_{10}, respectively 0.49 and 0.48 for LUR).

	NO_2	95%CI	PM_{10}	95%CI
SIRANE - NEAR	-0.10	-0.15 - -0.06	-0.09	-0.14 - -0.04
SIRANE - NEARMAIN	0.30	0.26 - 0.35	0.30	0.25 - 0.35
SIRANE - BUF150	0.29	0.24 - 0.34	0.24	0.2 - 0.29
LUR - NEARS	-0.12	-0.16 - -0.07	-0.13	-0.18 - -0.09
LUR - NEARMAIN	0.34	0.29 - 0.38	0.33	0.29 - 0.38
LUR - BUF150	0.33	0.28 - 0.37	0.31	0.26 - 0.36

Table 3.11: Cohen's kappa ($w\kappa$) between SIRANE/LUR and proximity models

Cohen's weighted Kappas, showed in table 3.11, roughly confirm the information given by the Spearman's rank coefficients. As for the comparison between the different models, data were categorized into quintiles to allow the calculation of $w\kappa$. With respect to the Viera and Garrett [2005], a fair agreement have been calculated, both for PM_{10} and NO_2, between the NEARMAIN metric and the models. A fair agreement, with slightly higher values of kappas, was also observed for the BUF150 metric.

Tables 3.12 and 3.13 show the results for the PA population:

	NO_2	PM_{10}		NO_2	PM_{10}
SIRANE - NEAR	0.19	0.19	LUR - NEAR	0.14	0.15
SIRANE - NEARMAIN	0.38	0.38	LUR - NEARMAIN	0.40	0.38
SIRANE - BUF150	0.42	0.41	LUR - BUF150	0.46	0.42

Table 3.12: Spearman's ρ between SIRANE/LUR and proximity models, PA population

	NO_2	95%CI	PM_{10}	95%CI
SIRANE - NEAR	0.11	0.08 - 0.13	0.12	0.09 - 0.14
SIRANE - NEARMAIN	0.24	0.21 - 0.26	0.24	0.22 - 0.27
SIRANE - BUF150	0.28	0.26 - 0.31	0.28	0.25 - 0.3
LUR - NEAR	0.10	0.08 - 0.13	0.10	0.07 - 0.12
LUR - NEARMAIN	0.26	0.23 - 0.28	0.26	0.23 - 0.28
LUR - BUF150	0.31	0.28 - 0.33	0.28	0.26 - 0.31

Table 3.13: Cohen's weighted Kappas $w\kappa$ between SIRANE/LUR and proximity models, PA population

Both for ρ and $w\kappa$, results are not much different than those referred to the real subjects. The main variations are due to a scale effect, remembering that the PA population counts 3000 subjects and the real one counts 785. That is the reason why the values for the LUR-NEAR and SIRANE-NEAR, who were negatives for the real population analyses, in this case are located in an expected range (very weak correlation but at least major than 1), both for the ρ and the $w\kappa$.

3.1.6 Multiple exposure evaluation

Another interest of the exposure assessment was to evaluate whether the subject evaluated as the most exposed to NO_2 are also the most exposed to PM_{10}, in order to evaluate the models' capability to describe multiple exposures.

The correlation between exposure to NO_2 and PM_{10} was calculated by the Pearson's r, Spearman's rank coefficient and the weighted inter-quintile kappa ($w\kappa$). Figure 3.8 shows scatterplots and linear regression R^2 between the two pollutant for SIRANE and LUR.

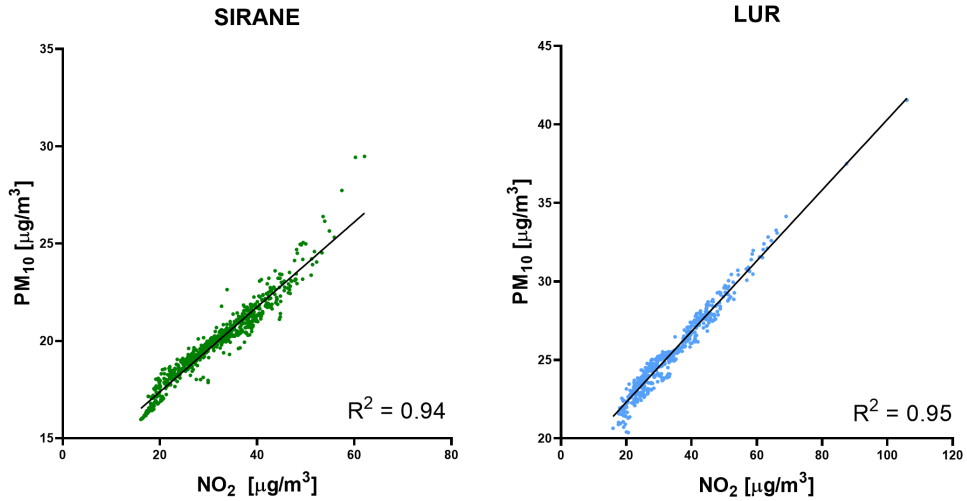

Figure 3.8: Scatterplots between NO_2 - PM $_{10}$ for SIRANE (left) and LUR (right) and linear regression lines

Pearson's r resulted extremely high for both models (0.98 for SIRANE, 0.97 for LUR). Spearman's ρ resulted almost equal to 1 (0.98 for SIRANE and 0.96 for the LUR model). Finally, $w\kappa$s indicate an almost perfect agreement level, with both models showing the same estimated value and confidence interval ($w\kappa$ = 0.86, CI: 0.84 - 0.88). It can be concluded that for both models subjects exposure to NO_2 and PM_{10} are ranked similarly and the exposure to the two pollutant are directly dependent following a linear proportionality (figure 3.8).

High values of correlation between exposure to NO_2 and PM_{10} are common in literature for different modeling approaches, which is expected considering they both derive from the same combustion sources (mainly urban traffic). Coudon et al. [2019b] observed a correlation of r = 0.9 between NO_2 and PM_{10} in a study involving 9 pollutants and more than 60000 real addresses (E3N cohort), using CHIMERE for the concentration estimations. Considering real measured values, correlation coefficients are lower: in a study on 31 Chinese cities, Xie et al. [2015] calculated a mean value of r between NO_2 and PM_{10} of 0.49.

3.1.7 Comparison within different land use type

The analysis of land use types no longer relies on the defined populations but to area values, according to the surfaces identified on the domain by the CORINE Land Cover database. The mean and median values of SIRANE and LUR cells were calculated in the different geometries present within the domain. The analyses have been performed both for the NO_2 and PM_{10}, but since PM_{10} does not exhibit any pattern (LUR overestimates SIRANE's values not depending on the land use type considered), only results for the NO_2 are presented.

Ozone was not included into the analysis because the interpretation of the results would have been extremely difficult, because the spatial resolution is not sufficient to permit a valid discussion.

Figure 3.9 shows the mean and median value for SIRANE and LUR cells for the CLC level 1.

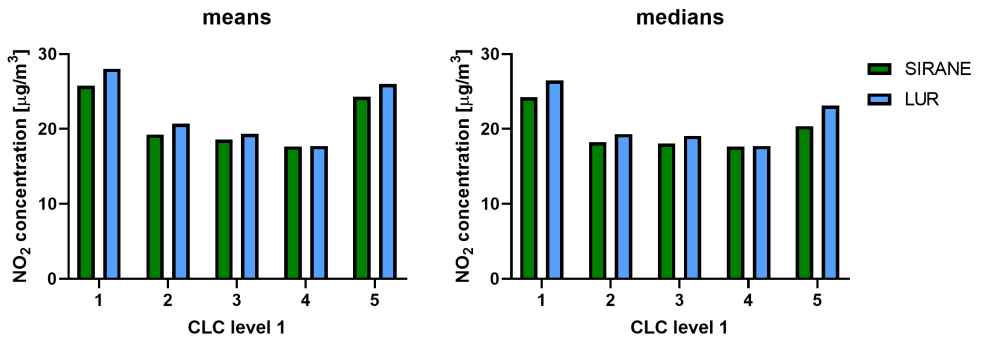

Figure 3.9: Means and medians for NO_2 values through the geometries of CLC at level 1. The complete CLC nomenclature is presented in Appendix D.

The class that shows the highest average and median exposition is 1 (artificial surfaces), as expected. The high values for the class 5 (water bodies) are certainly linked to the rivers that flow very closely to the city center and to the industrial areas in the southern part of the domain, usually bordered by the major traffic routes of the city. Averaged values estimated by

LUR are generally higher, especially in the sector 1 (28.04 $\mu g/m^3$ vs 25.77 $\mu g/m^3$). Furthermore, the differences between means and medians are not significant (averagely around 1.20 $\mu g/m^3$), indicating that both for LUR and SIRANE the calculation of mean values within the CLC geometries is not strongly affected by peaks of concentration. The major difference between means and medians is observed for the class 5 (water bodies), whose distribution is characterized by extremely high values on the geometries' sides (resulting from the influence of the bordering roads) and extremely low values on their central areas (of course due to the fact that into the rivers there are not NO_2 emission sources). For SIRANE the difference between mean and median of sector 5 is equal to nearly 4 $\mu g/m^3$, while for LUR is 2.90 $\mu g/m^3$.

Figure 3.10 shows the prosecution of the analysis at CLC level 2:

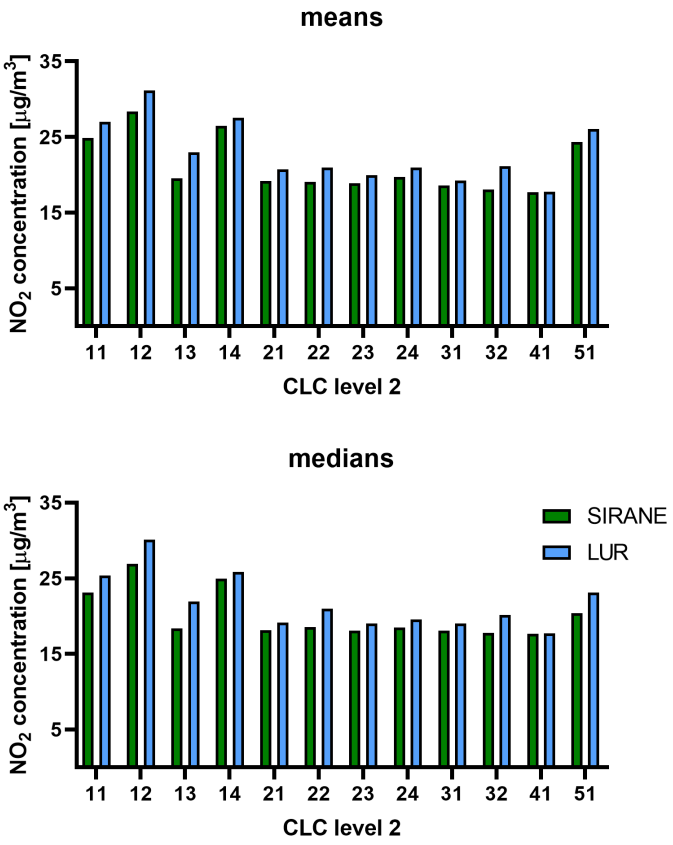

Figure 3.10: Means and medians for NO_2 values through the geometries of CLC at level 2

It is observable (figure 3.10) that the high values related to the sector 1 are attributable to the subsectors 11, 12 and 14:

This is quite expected, being the land use types usually associated with the major presence of emission sources both linear (traffic roads), punctual (chimneys and other industrial exhausts) and areal (industrial sites). The other CLC code with high emissions observed is 51, which is continental waters and again refers to rivers, for which the considerations already made are valid.

LUR confirms to estimate values averagely higher than SIRANE: in term of means, the subsectors where the difference is larger are the 12

CODE	NAME	SIRANE	LUR
11	Urban fabric	24,85	27,00
12	Industrial, commercial and transport units	28,36	31,13
14	Artificial non-agricultural vegetated areas	26,44	27,54

(Industrial, commercial and transport units) and the 13 (Mine, dump and construction sites). Median values confirm not to be distant to means: in this characteristic, the resolution of the model played a key role. In fact, for SIRANE (whose resolution is 10 meters) the mean value is averagely 1.30 $\mu g/m^3$ higher than the median, while for LUR the difference is lower (0.8 $\mu g/m^3$).

Figure 3.11 shows the analysis performed at finals CLC codes, focusing on class 1 because of major interest:

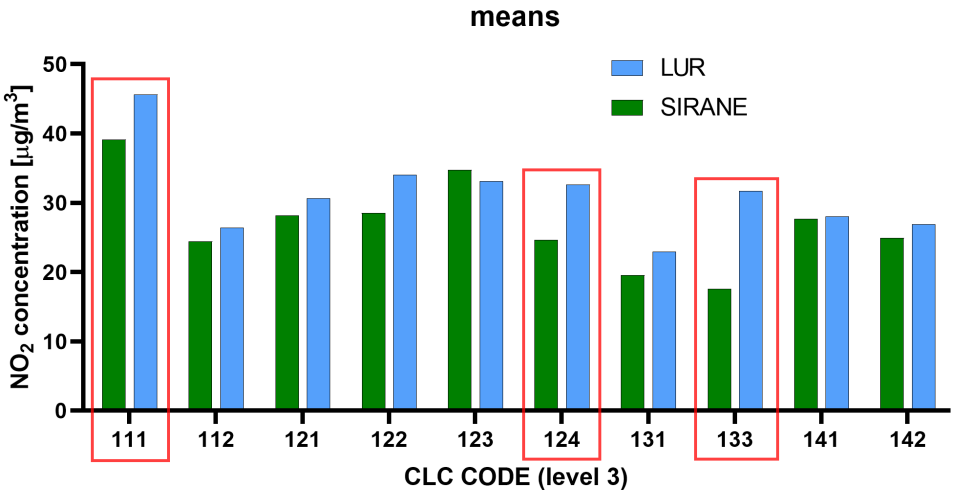

Figure 3.11: Means and medians for NO_2 values through the geometries of CLC at level 3, focusing only on class 1

The highlighted CLC CODEs in figure 3.11 are those where the highest difference has been observed:

CODE	NAME	SIRANE	LUR
111	Continuous urban fabric	39,12	45,63
124	Airports	24,61	32,64
133	Construction sites	17,56	31,70

The extremely big variations in sectors 124 (Airports) and 133 (Construction sites, where LUR is more than 80% higher) probably depend on very small-scale variations that LUR is unable to capture. Anyway, these two subsectors refer to very limited areas within the domain (map in Appendix D).

The CLC sector resulting as the most exposed to NO_2 was the continuous urban fabric (111), on which both models were capable to identify a substantial difference, namely from other subsectors within the urban fabric and the industrial units' level. However, in sector 111 LUR significantly overestimate SIRANE (45,63 $\mu g/m^3$ vs 39,12 $\mu g/m^3$, 17% higher): this is probably due to the fact that for the LUR model a high density of civil buildings is considered as a strong predictor for NO_2 concentrations, as explained in section 2.4.

3.1.8 Comparison within different average income groups

The association between air pollution (mainly NO_2 and PM_{10}) and socioeconomic factors, such as average income or social deprivation indices, has dealt with in several [Deguen and Zmirou-Navier, 2010] and is a common application of air pollution models. The objective of the analysis was to investigate whether SIRANE and the LUR models show evident differences in evaluating the potential association between a socio-economic factor and the exposure to PM_{10} and NO_2, qualitatively assessing their interchangeability for those studies. Data about the average income were available within the domain at the IRIS level, as explained in section 2.8, finally resulting in 609 paired values of average income and pollutant concentration.

Average incomes were categorized into quintiles, from 1 (20% IRIS with lower income) to 5 (20% higher), and data about pollutant concentrations were split into 5 groups depending on those.

NO_2

Figure 3.12 shows the distribution of NO_2 values for each income group, both for SIRANE and LUR.

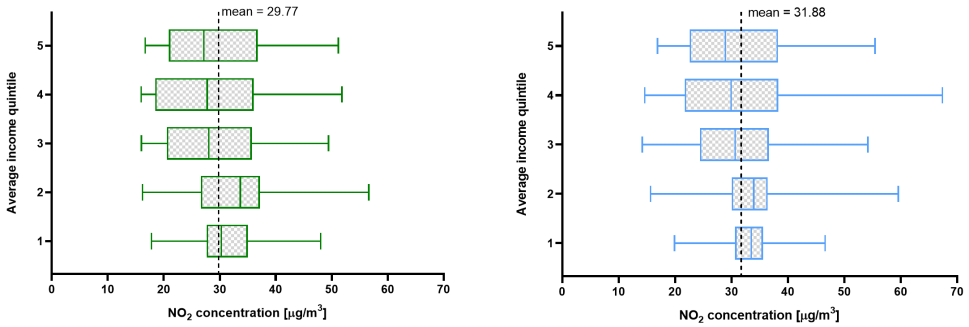

Figure 3.12: Inter quintile distribution of NO_2 average values at IRIS level

The two models present a comparable inter-group variability, with average values that are highest for the first group and progressively slightly

	SIRANE		LUR	
Income quintile	mean	sd	mean	sd
1	31.29	6.14	32.97	4.71
2	31.67	7.35	33.40	7.35
3	28.51	8.39	30.59	8.58
4	28.40	9.50	31.04	11.01
5	28.98	8.65	31.41	10.53

Table 3.14: Mean and standard values per income quintile group, NO_2 [$\mu g/m^3$]

decrease. An interesting result is also the one regarding the standard deviation, since for both models the data variability increases considering higher income quintiles. A similar result was observed by Morelli et al. [2019] in a study about social deprivation and exposure to $PM_{2.5}$ in Grenoble and Lyon. One possible interpretation of this is that high income people live both in the hilly areas outside of the city (which of course suffer less pollution) and within the old town (which is very polluted due to the presence of street canyons and intense traffic). On the contrary, the low incomes are concentrated in suburban neighborhoods, which averagely suffer high pollution levels.

The statistical difference between exposures among income quintiles has been assessed through the non-parametric Wilcoxon test, because some of the distributions could not be considered as Gaussian. The difference has been tested between selected pairs of groups (1-4, 1-5, 2-4 and 2-5): the test was "one sided", i.e. that the alternative hypothesis was exposure for groups 4-5 was inferior to those for groups 1-2. The null hypothesis of the test, that was the equality of the means, was always rejected at 95% confidence (table 3.15):

It is observed that both SIRANE and LUR identify a significant difference between exposure in all the analyzed quintile pairs, with groups 4-5 being significantly less exposed than groups 1-2.

quintiles	p-value SIRANE	p-value LUR
1-4	0.004	0.004
1-5	0.004	0.001
2-4	0.002	0.004
2-5	0.006	0.004

Table 3.15: p-value resulting from Wilcoxon test between data in different quintiles. A p-value lower than 0.05 indicate that the means of the two distribution cannot be assumed as equal at 95% of significance [Kottegoda and Rosso, 2008]

PM_{10}

Figure 3.13 and table 3.16 summarize the results obtained for PM_{10}:

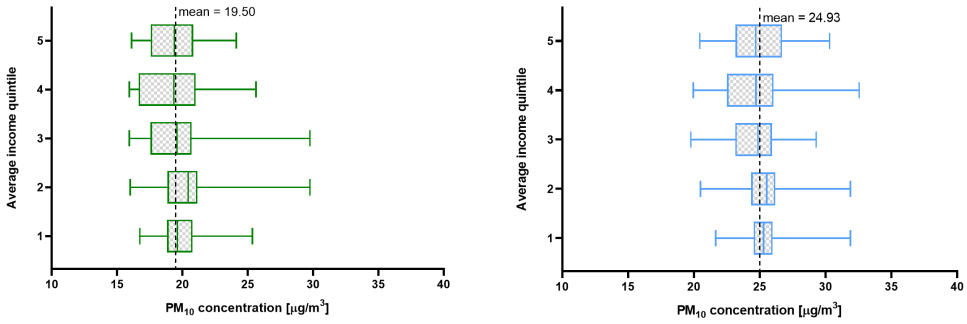

Figure 3.13: Inter quintile distribution of PM_{10} average values at IRIS level

For the PM_{10}, it is observable that the increasing variability showed by the NO_2 with income quintile is less accentuated, especially for LUR. p-values resulting from the Wilcoxon nonparametric test are showed in table 3.17:

The p-values are generally higher. SIRANE keeps individuating all p-values well lower than 0.05. On the contrary, the LUR model does not capture the 95% significant difference between 1-5 (p-value 0.057) and barely identifies the one between 2-5 (p-value 0.048)

	SIRANE		LUR	
Income quintile	mean	sd	mean	sd
1	19.91	1.62	25.19	1.36
2	20.04	1.89	25.33	1.91
3	19.23	2.18	24.57	2.10
4	19.15	2.30	24.66	2.57
5	19.31	1.92	25.05	2.30

Table 3.16: Mean and standard values per income quintile group, PM_{10} [$\mu g/m^3$]

quintiles	p-value SIRANE	p-value LUR
1-4	0.009	0.025
1-5	0.013	0.057
2-4	0.002	0.008
2-5	0.003	0.047

Table 3.17: p-value resulting from Wilcoxon test between data in different income quintile group

.

Differences

A further interest of the investigation is in verify whether SIRANE and LUR tend to show differences in estimating pollutant concentration (and consequently in assigning exposures) depending on the average income quintile. Table 3.18 shows the average difference value between SIRANE and LUR:

Even if LUR demonstrate to overestimate SIRANE in each group for both pollutants, which was expected, values of differences for higher average income are greater, especially for the fourth and fifth group.

The nonparametric Wilcoxon test was newly performed to quantify if the average SIRANE-LUR difference statistically vary in function of the group.

For the NO_2 distributions, there is not a statistical evidence that the

Income quintile	Average difference (NO_2)	Average difference (PM_{10})
1	-1.68	-5.28
2	-1.73	-5.29
3	-2.08	-5.34
4	-2.63	-5.51
5	-2.43	-5.73

Table 3.18: Average difference SIRANE - LUR for income quintile group, [$\mu g/m^3$]

quintiles	p-value NO_2	p-value PM_{10}
1-4	0.36	0.25
1-5	0.46	$5.00 \cdot 10^{-4}$
2-4	0.18	0.08
2-5	0.22	$1.62 \cdot 10^{-5}$

Table 3.19: p-values resulting from a Wilcoxon test for the SIRANE-LUR difference distribution between 1-4, 1-5, 2-4 and 2-5 quintile paired groups

SIRANE-LUR differences distribution's mean vary depending on average income group. On the other hand, PM_{10} difference distributions' means were evaluated as significantly different for the 2-5 and 1-5 pairs. This means that the average difference between SIRANE-estimated and LUR-estimated values is statistically different between group 1 and 5 and between group 2 and 5. It is observed that the two models tend to have a higher disagreement among them in estimating PM_{10} exposure for high-income people compared to low-income ones. Anyway, they both show a capacity to identify a clear decrease in NO_2 exposure and an increase in data variability for higher average income groups with respect to the lowers.

3.2 Year 2000

The comparison of the models for the year 2000 has been carried out just between the model SIRANE and LUR. Considering the analysis in 2010, these two models showed good correlation in predicting the subjects' exposure to atmospheric pollution (see section 3.1.3). Correlation and inter-rater agreement were evaluated through the same methodology used for 2010. Differences between the two datasets were also identified and the evolution of the exposure values estimated by the two models from 2000 to 2010 was assessed.

3.2.1 Data description and graphics

The summary statistics are given in table 3.20, both for SIRANE and LUR and for three pollutants, while boxplots are shown in figure 3.14.

	NO_2		O_3		PM_{10}	
	SIRANE	LUR	SIRANE	LUR	SIRANE	LUR
Min	23.46	20.57	12.10	42.50	19.93	25.57
1^{st} Quartile	34.78	37.65	28.67	43.82	23.37	32.68
Mean	44.13	44.70	34.36	45.34	25.78	34.17
Median	44.96	44.67	33.26	45.45	25.55	34.58
3^{rd} Quartile	52.68	49.93	39.99	46.23	27.47	35.41
95^{th} Quantile	62.64	63.56	46.61	48.76	31.47	38.69
Max	90.52	119.05	48.17	51.52	46.25	51.37
SD	11.75	11.26	7.11	1.86	3.62	3.13

Table 3.20: Summary statistics for the real population in 2000; all data are in $\mu g/m^3$

For the NO_2, the standard deviation is similar, around 11 $\mu g/m^3$. Means and medians are both around 44 $\mu g/m^3$, while SIRANE has a higher IQR (17.90 vs 12.28 $\mu g/m^3$), indicating that for SIRANE NO_2 is more heterogeneously distributed over the domain. Regarding high val-

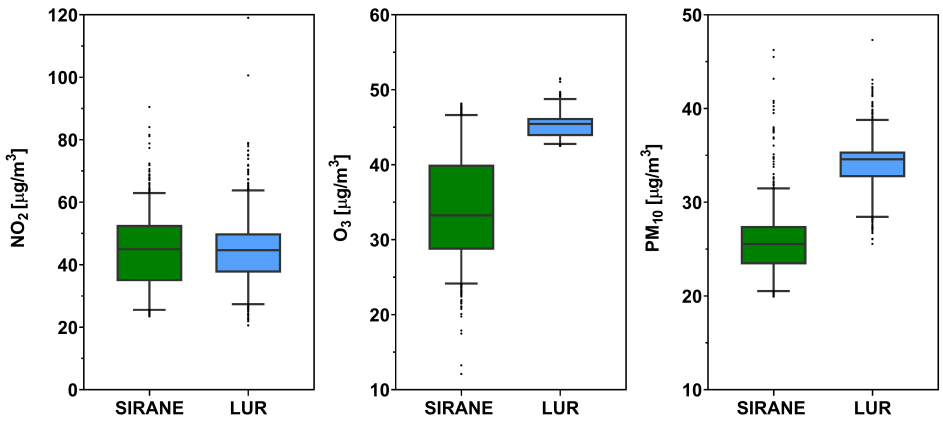

Figure 3.14: Boxplots for the real population in 2000

ues, even if 95th quantiles are quite close (62.64 and 63.56 $\mu g/m^3$) the LUR provides a greater maximum value (around 120 vs 90 $\mu g/m^3$). The percentage of subjects exposed over the limit provided by the WHO guidelines for the annual average exposure (40 $\mu g/m^3$) is 61.91 % for SIRANE and 68.91% for LUR (in 2010 they were both under 20%) [WHO, 2005].

The O_3 situation is very different, as the two datasets have very different SD (7.11 $\mu g/m^3$ for SIRANE, 1.86 $\mu g/m^3$ for LUR) and have different central values (the minimum value for LUR, 42.50 $\mu g/m^3$, is higher than the 75th quantile of SIRANE). As in 2010 (see table 3.1 and figure 3.1), the LUR model provides data which have a very low capability of capturing intra-urban variations (see also the LUR ozone map in Appendix C). Indeed, all LUR values are set between 42.50 and 51.52 $\mu g/m^3$.

For the PM_{10}, the two SD are similar, even though SIRANE's is a little higher (3.62 vs 3.13 $\mu g/m^3$). Furthermore, considering figure 3.14, LUR generally overestimate SIRANE values (LUR's first quartile is equal to 32.69 $\mu g/m^3$, higher than SIRANE's 95th quantile that is 31.47 $\mu g/m^3$). The ratio between the 99th and 95th quantile is 1.26 for SIRANE and 1.08 for LUR, indicating that the dispersion model provides a more accurate

description of concentration peaks within the city. Both for SIRANE and LUR, all the subjects of the real population are exposed to concentrations above the limits of the WHO air quality guidelines for the annual average exposure (20 $\mu g/m^3$).

Histograms shown in figure 3.15 have similar shape than those referring to the year 2010 (figure 3.2).

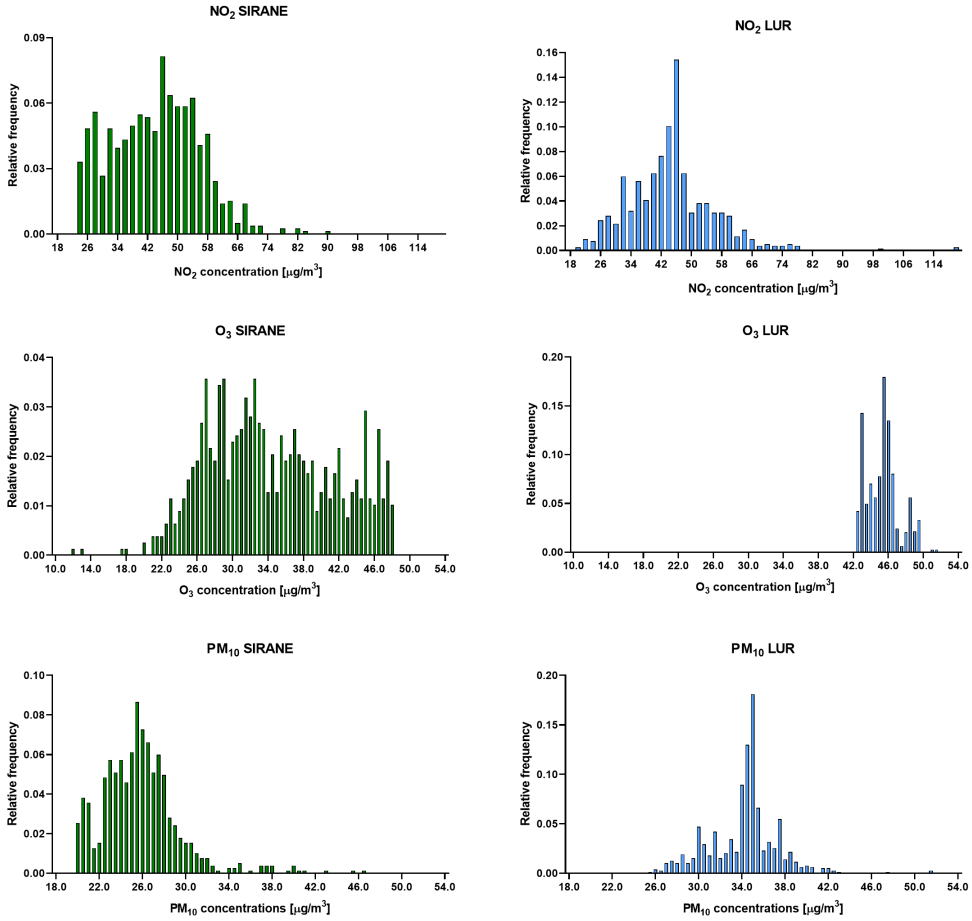

Figure 3.15: Histogram representations for the real population in 2000

NO_2 and PM_{10} are both for LUR and SIRANE heavy-tailed on the right, having lots of estimated values far from the "head"of the distribution and corresponding to highly polluted locations within the domain (major roads, industrial areas).

Figure 3.16 shows scatterplots and linear regression lines between the two models for three pollutant.

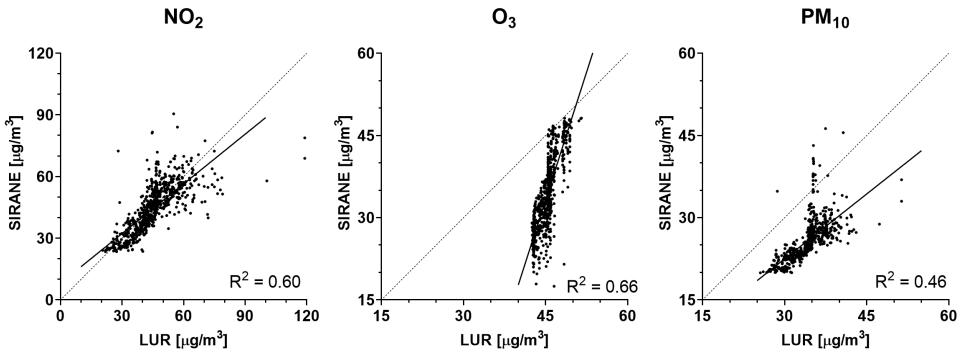

Figure 3.16: Scatterplot and linear regression lines for the real population in 2000

Scatterplots are similar to those in figure 3.4. A few points seem to be over the bisector regarding the O_3, meaning that almost the totality of subjects is more exposed to O_3 in LUR than in SIRANE, and the same can be said for the PM_{10}. R^2 values indicate that good amounts of SIRANE variability can be predicted by the LUR model through linear regression, especially for ozone (almost 66%).

3.2.2 Differences and similarities with PA population

The PA population datasets are summarized in table 3.21.

	NO_2		O_3		PM_{10}	
	SIRANE	LUR	SIRANE	LUR	SIRANE	LUR
Min	24.09	22.65	10.28	42.46	20.05	25.76
1^{st} Quartile	33.09	35.96	32.13	45.17	23.11	31.53
Mean	40.24	42.35	36.64	45.92	24.73	33.48
Median	38.16	42.01	37.69	45.86	24.09	33.88
3^{rd} Quartile	46.59	46.63	41.19	46.64	25.82	34.85
95^{th} Quantile	56.10	58.49	45.12	48.79	28.60	37.72
Max	151.07	94.19	47.69	51.49	57.28	45.84
SD	9.81	9.58	5.90	1.67	2.86	2.76

Table 3.21: Data description for the PA population in 2000; all data are in $\mu g/m^3$

All data have a minor SD, that is expected considering that the numerosity of the population increased from 785 to 3000 subjects. The NO_2 shows a minor IQR for SIRANE, becoming less far from LUR's (13.50 and 10.68 $\mu g/m^3$). Another difference refers to great exposure values, where SIRANE presents a very high maximum (151.07 $\mu g/m^3$) that is due to the characteristic of the PA population shapefile already described at the beginning of section 3.1.2.

Histograms (figure 3.17) presents heavy-tailed shapes for the NO_2 and PM_{10} due to the higher numerosity of the population, showing also a more defined clustering of values in the left part of the graph. The bimodal behavior of the O_3 in the SIRANE simulation, just as already seen in 2010 (figure 3.6), is also highlighted.

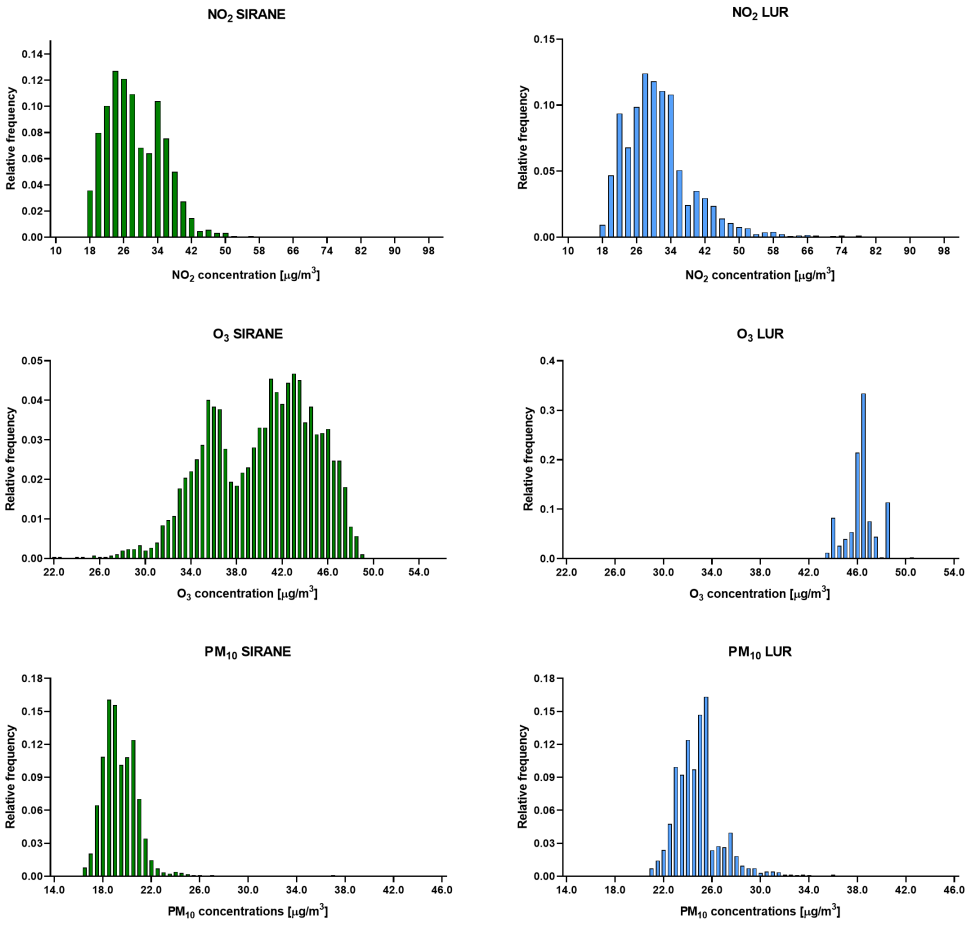

Figure 3.17: Histogram representation for the PA population in 2000

3.2.3 Correlation and agreement coefficients

The correlation and agreement level between SIRANE and LUR were evaluated through the same statistical indicators calculated for 2010 in section 3.1.3. This further analysis is meant to verify that the good performance of these two models found for 2010 are extendable to the year 2000, which is what is expected.

Table 3.22 shows Pearson's r and Spearman's ρ for the three pollutants:

Generally, a good correlation was observed. Pearson's r values are higher than those found in 2010, except for PM_{10} (0.72 in 2010, 0.68 in

Indicator	NO_2	O_3	PM_{10}
Pearson's r	0.77	0.81	0.68
Spearman's ρ	0.86	0.83	0.86

Table 3.22: Correlation coefficients between SIRANE and LUR in 2000, real population

2000). On the contrary, Spearman's ρ indicate better rank-correlation for all pollutants, with high values both for NO_2 and PM_{10} (ρ = 0.86).

Table 3.23 shows Cohen's $w\kappa$ calculated for the real population in the year 2000, referring to inter-quintile agreement:

	NO_2	O_3	PM_{10}
Cohen's wκ	0.67 (0.64 - 0.71)	0.63 (0.59 - 0.66)	0.67 (0.64 - 0.70)

Table 3.23: Cohen's wκ for the real population in 2000. Values between parentheses indicate the 95% confidence interval.

Colours in table 3.23 refers to the classification made by Viera and Garrett [2005] and indicate that for the three pollutant there is a "substantial agreement"between LUR and SIRANE, confirming in the year 2000 the results already observed for 2010.

3.2.4 Comparison between data for year 2010 and year 2000

Data for the years 2000 and 2010 have also been compared with the aim of defining the average differences between two endpoints of a 10-year period and further investigate the capability of the two models of retrace past exposition trajectories.

Figure 3.18 shows a boxplot comparison for NO_2:

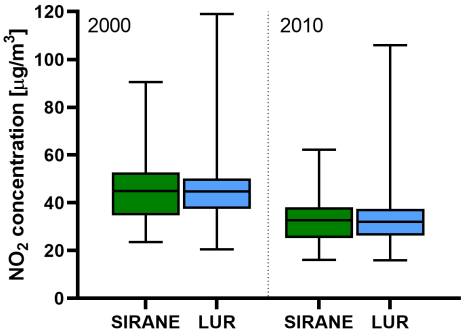

Figure 3.18: NO_2 for SIRANE and LUR, comparison between 2000 and 2010

As already seen in previous chapters, LUR distributions generally shows greater high tails, while SIRANE's IQRs are slightly higher. It is observable that the exposure values have decreased with time. Another evidence is that there is no apparent difference in the way SIRANE and LUR represent this evolution, since both in 2000 and in 2010 they are almost placed into the same data range. Furthermore, for both models the boxplots for 2010 have a more flattened shape than those for 2000.

Some considerations are valid also for PM_{10} and O_3: as it is possible to see in figure 3.19, boxplots for the year 2010 suggest that the differences between the two models are similar in 2010 and in 2000. This was expected since the data description and the calculation of the correlation and agreement coefficients for the paired distributions both in 2000 and in 2010 showed that there was no substantial difference in their relative behavior between the two years.

Table 3.24 shows the average differences between 2010 and 2000, with

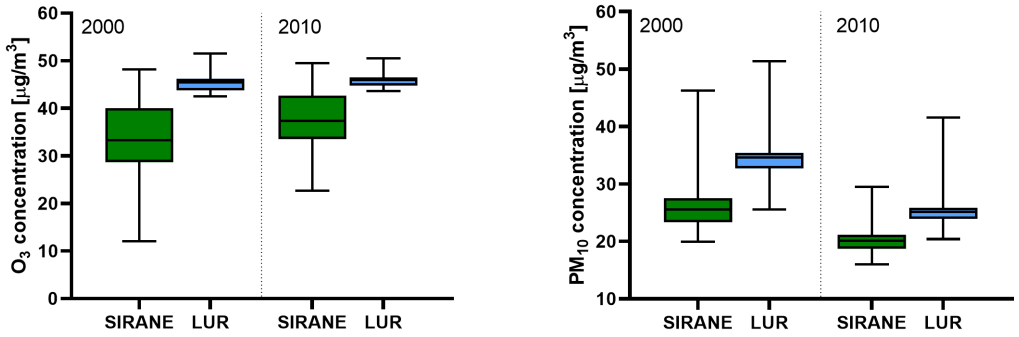

Figure 3.19: O_3 and PM_{10} for SIRANE and LUR, comparison between 2000 and 2010

their relative confidence intervals at 95%:

	NO_2	95%CI	O_3	95%CI	PM_{10}	95%CI
SIRANE	-12.16	(-11.92 ; -12.40)	3.75	(3.87 ; 3.64)	-5.46	(-5.38 ; -5.54)
LUR	-12.17	(-11.71 ; -12.34)	0.62	(0.66 ; 0.58)	-9.14	(-9.01 ; -9.31)

Table 3.24: Median differences between values for the year 2000 and 2010 for the real population, values in $\mu g/m^3$. 95% CI were provided by a Wilcoxon test

For the NO_2, a clear decrease in average values from 2000 to 2010 is observed both for SIRANE and LUR (around -12 $\mu g/m^3$ for both models). It is noticeable that the LUR model has greatly reproduced SIRANE average variation throughout a 10-year period for the estimation of the exposure to NO_2 of the real population. The decrease is easily attributable to the lowering of the traffic emission intensity, because of both increasingly careful policy strategies and of the partial evolution of the urban car fleet towards low-emission technologies, above all the fitting of exhaust after-treatment systems in diesel vehicles. The decrease indicated by both models is in agreement with the recent trends in NO_2 concentrations in Europe, especially in urban environment: Colette et al. [2011] indicate for the period 1998-2007 a very strong decrease observed for NO_2 in urban air

quality monitoring stations in Europe, with a median decreasing trend of -0.37 $\mu g/m^3$ per year.

An important decrease from 2000 to 2010 was observed also in PM_{10} both for SIRANE and LUR, respectively equal to 5.46 $\mu g/m^3$ and 9.14 $\mu g/m^3$, with the LUR individuating a clearly higher difference. The decreasing trend for the PM_{10} concentrations is consistent with observed trends in monitoring stations: Guerreiro et al. [2014] individuated in France for the period 2002-2011 a negative trend of averagely -0.56 $\mu g/m^3$ per year considering urban background monitoring stations and -0.74 $\mu g/m^3$ per year considering traffic-related measures.

A behavior opposite to that of NO_2 and PM_{10} is the ozone's, the only pollutant whose concentrations increased during the considered period. SIRANE estimated an average increase of 3.75 $\mu g/m^3$ over the real population's subjects. The LUR estimates a lower increase (0.64 $\mu g/m^3$), probably because of its lower heterogeneity within the domain. A study describes the ozone trends over all France in the period 1999-2012, showing that mean concentrations increased for the 66.2% of measurement stations over France (76.5% if considering only urban sites)[Sicard et al., 2016]. Several causes could have led to an improvement of ozone concentrations: one of the reasons could be that, during the cold period, the increase in O_3 mean concentrations in urban stations can be attributed to the lower effect of the O_3 titration by NO as a consequence of the reduced NO_x emissions trends within European countries [Doherty et al., 2005]. Furthermore, during summers ozone concentrations increases due to higher temperatures and reduced cloudiness and precipitation over Europe as a consequence of the climate change [Meleux et al., 2007].

Figure 3.20 shows scatterplots and linear regression lines for all pollutants' estimations by SIRANE and LUR between the year 2000 and 2010. NO_2 and O_3 values for SIRANE are quite aligned with the regression line, meaning that there is a great proportionality between how low and high values decreased from 2000 to 2010. For NO_2, the major decrease is observed in high values: for example, 75th quantile decreased from 52.68 $\mu g/m^3$ in 2000 to 38.13 $\mu g/m^3$ in 2010 (-14.55 $\mu g/m^3$), while the 95th from

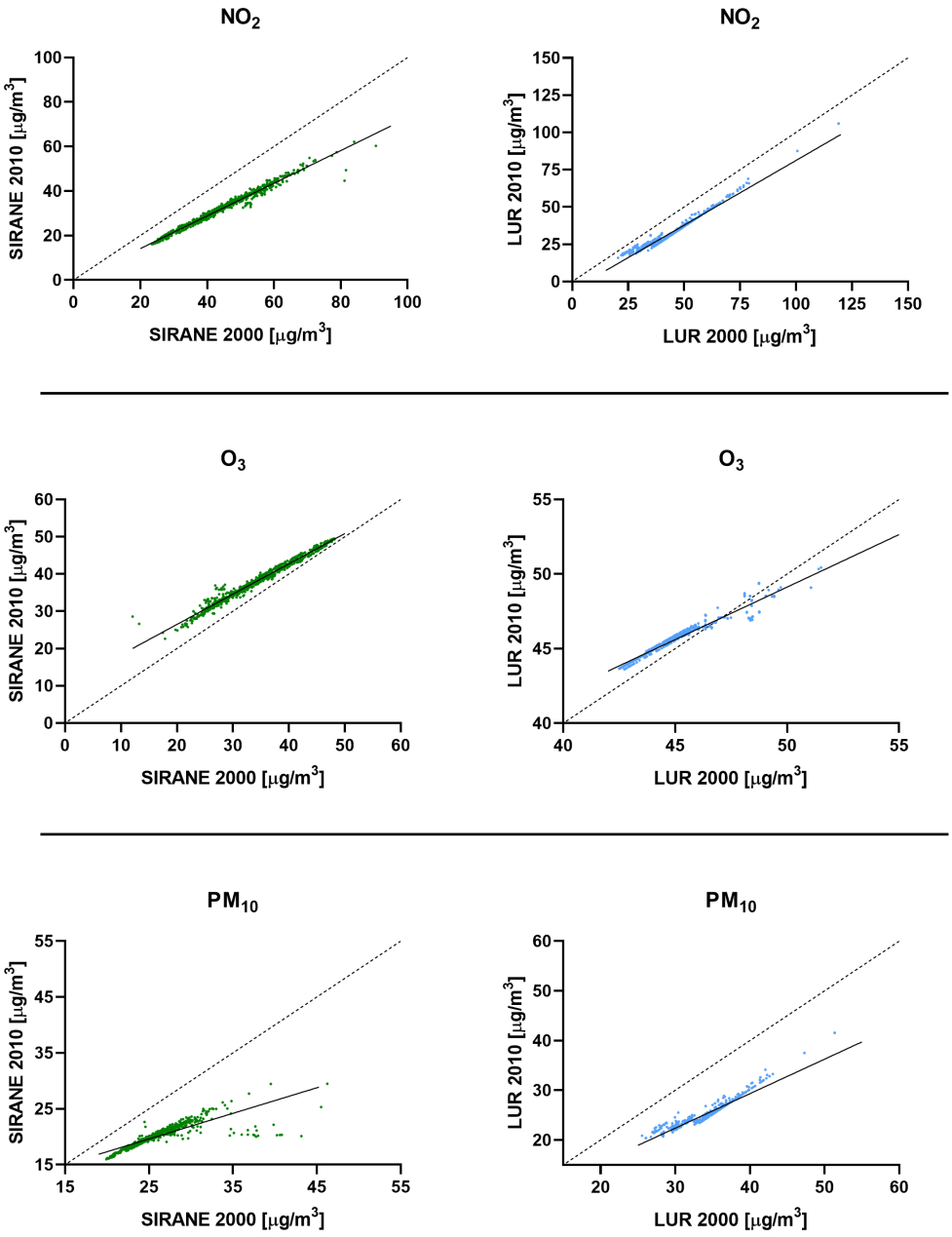

Figure 3.20: Scatter-plots of NO_2, O_3 and PM_{10} concentration between 2000 and 2010 for LUR and SIRANE

62.64 $\mu g/m^3$ to 45.67 $\mu g/m^3$ (-16.97 $\mu g/m^3$). It can be concluded that value have decreased proportionally, while for the LUR this is less accentuated. On the contrary, ozone values that increased most are the central ones: while 95th quantile are almost the same between the two years (around 46 $\mu g/m^3$, the median improved noticeably (33.26 $\mu g/m^3$ in 2000, 37.32 $\mu g/m^3$ in 2010).

	SIRANE			LUR		
pollutant	intercept	slope	R^2	intercept	slope	R^2
NO_2	-0,55	0,73	0,98	-5,64	0,87	0,97
O_3	10,21	0,81	0,98	13,93	0,70	0,95
PM_{10}	8,19	0,46	0,71	1,59	0,69	0,90

Table 3.25: Linear regression coefficients for paired 2000-2010 estimated exposures

The only prediction with a R^2 inferior to 0.9 (0.71) is the PM_{10} for SIRANE, having a linear regression line that most of all differs from the bisector. PM_{10} shows the same property of the NO_2 (the proportionality between dimension and decrease), with about twenty points that underwent a very strong decrease (clearly visible in figure 3.20). Those points represent subjects located in a specific neighborhood within the city (*Tonkin Sud*, in the *VI arrondissement*) that has seen a huge decrease in PM_{10} concentration due to the shutdown of an important PM_{10} stationary source (a cogeneration plant). This characteristic is not completely captured by the LUR model, whose points are well aligned with the linear regression line (R^2 = 0.90).

3.3 Odds ratio calculus comparison

Final aim of the exposure assessment is the estimation of epidemiological outputs, i.e. statistical indicators used to assess whether a disease occurrence is linked or not with the studied factor (for example tobacco, air pollution, drugs). For case-control studies, odds ratios measure how much an adverse health effect occurs in highly exposed subjects with respect to the same effect occurrence observed in less exposed ones. In the framework of this study, the disease considered is breast cancer. The exposure to high levels of air pollution, with respect to previous finding in epidemiological studies, may increase the risk of breast cancer of 5 to 15% [White et al., 2018]. The relative capability of SIRANE and LUR of estimating inter-quartile odds ratios (i.e. between 4 exposure groups, Q1 to Q4, defined by the distribution quartiles) have been compared through 500 random population of 10000 subjects over the domain, regarding their exposure to NO_2 and PM_{10} for the year 2010. For each one of the 500 populations, the spatial distribution of cases and controls was defined so that inter-quartile odds ratio values for SIRANE were equal to a reference, showed in figure 3.26, and then LUR's odds ratios have been obtained replacing the exposure values (and so modifying the distribution of cases and controls through the 4 groups) with LUR's.

	Cases	Controls	Total	OR	CI95%
Q4	1364	1250	2614	1.20	1.07 - 1.34
Q3	1280	1250	2530	1.13	1.01 - 1.26
Q2	1220	1250	2470	1.07	0.96 - 1.2
Q1	1136	1250	2386	1.00	-
Total	5000	5000	10000		

Table 3.26: Cases and controls reference repartition

Figure 3.21 shows the spatial division of the study domain in exposition quartiles both for SIRANE and LUR, identifying the areas corresponding to the four exposure groups (Q1, Q2, Q3, Q4). The quartiles values

were calculated as weighted in function of the population density using the function “spatstat::weighted.quantile”on RStudio.

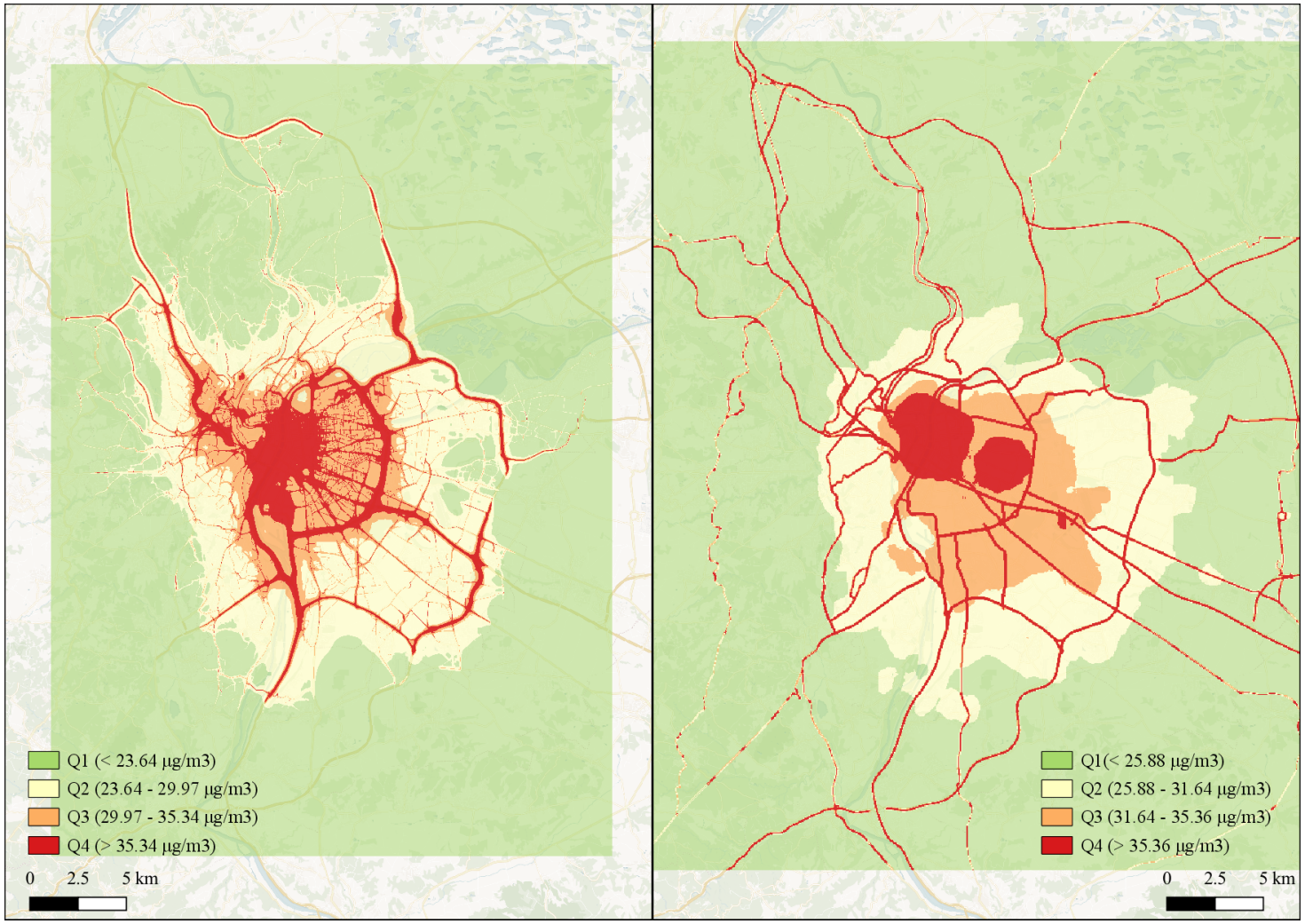

Figure 3.21: Spatial division of the domain in function of the density-weighted quartiles for NO_2, for SIRANE (on the left) and LUR (on the right)

Base scenario

The average odds ratios resulting from the 500 simulation for the NO_2 are presented in figure 3.22 and in table 3.27. In figure 3.22, also the maximum minimum values for the interval extremes within the 500 simulations are indicated.

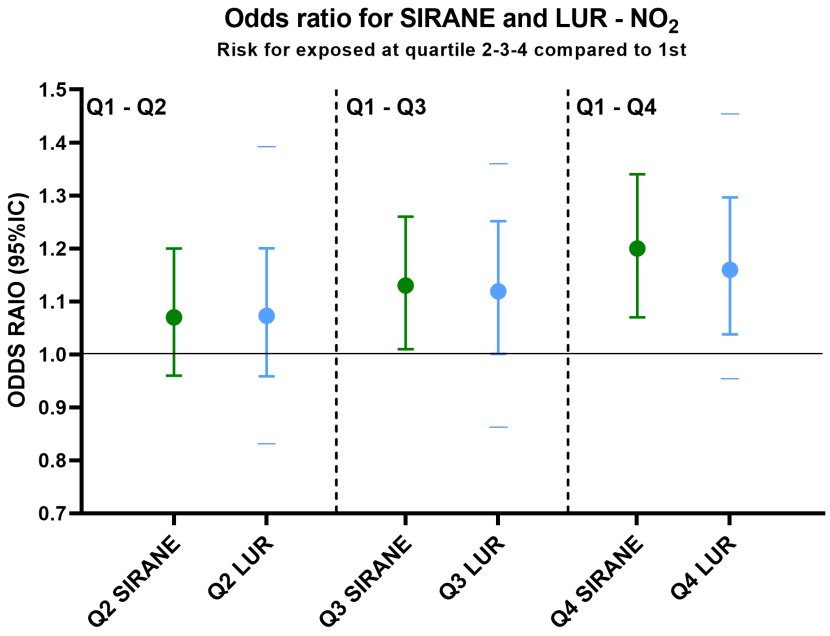

Figure 3.22: Inter-quartile odds ratios for SIRANE (reference, in green) and for LUR (blue); extreme values for the 95%CI in LUR are averaged between the 500 extreme values obtained, while their maximum and minimum are displayed as the isolated tracts

The odds ratio between the first and second quartile occupies the same interval of the reference (1.07, 95%CI: 0.96 - 1.20), while for the third and fourth quartile the LUR values are slightly inferior to SIRANE's.

Values set from table 3.26 (green points in figure 3.22) indicate that a statistically significant odds ratio is averagely observed for SIRANE between quartiles 1-3 and 1-4. Figure 3.23 shows the percentage of LUR simulations that keep individuating the significance:

	min	min (mean)	estimated	max (mean)	max
Q1	-	-	1	-	-
Q2	0.83	0.96	1.07	1.20	1.39
Q3	0.87	1.00	1.12	1.25	1.37
Q4	0.94	1.04	1.16	1.30	1.48

Table 3.27: Inter-quartile odds ratios estimated by LUR for the NO_2 exposure; min and max refer to 95% Confidence Intervals extremes

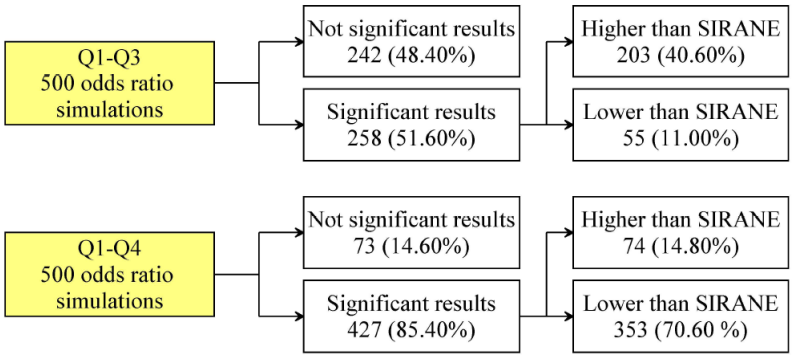

Figure 3.23: Percentage of epidemiologically significant odds ratio individuated by LUR over 500 simulations for NO_2

It is observed that around half of the simulations lost the significance for the inter-quartile odds ratio between the first and third quartile, of which the 40.60% estimate a value higher than SIRANE. For the odds ratio between first and fourth quartile, even if most of estimated values are lower than SIRANE's, more than 80% of simulations keep capturing a significant association.

Figure 3.24 and table 3.28 shows the results for PM_{10}:

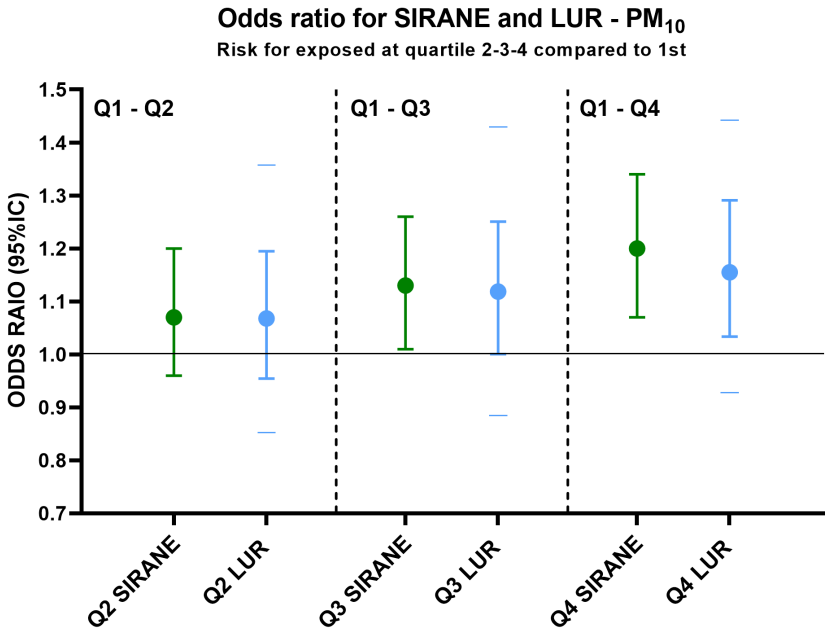

Figure 3.24: Inter-quartile odds ratios for SIRANE (fixed, in green) and for LUR (blue) for PM_{10}

	min	min (mean)	estimated	max (mean)	max
Q1	-	-	1	-	-
Q2	0.85	0.95	1.07	1.19	1.36
Q3	0.88	1.00	1.12	1.25	1.43
Q4	0.93	1.03	1.15	1.29	1.44

Table 3.28: Inter-quartile odds ratios estimated by LUR for the PM_{10} exposure

Results for PM_{10} were similar to those of NO_2, since while the odds ratio between quartiles 1-2 is almost equal to the reference, a little decrease is observed for quartiles 1-3 and 1-4 with respect to SIRANE. Percentages of simulations conserving the odds ratio's significance (figure 3.25) is also constant with NO_2 results.

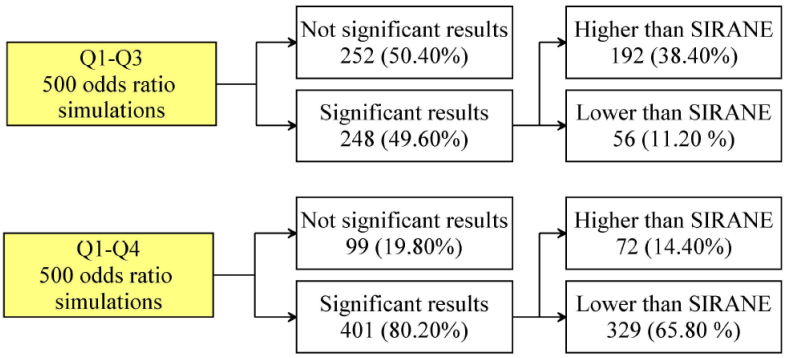

Figure 3.25: Percentage of epidemiologically significant odds ratio individuated by LUR over 500 simulations for PM_{10}

Highly exposed populations

The second scenario ("high exposure") was defined to assess model's performances between highly exposed subjects, defining an inferior threshold for the quartiles' calculation. Therefore, population density-weighted quartiles describe an area that is quite lower than those of the base scenario (illustrated in figure 3.26 and mainly constituted by the city center) and are based on smaller-scale variations.

The procedure applied was the same as for the base scenario: the only difference refers to the initial dataset from which the populations were sampled. For the NO_2, this dataset considered only cells where concentration values estimated by SIRANE were higher than 25 $\mu g/m^3$(around 2.5 million cells). The threshold value for the PM_{10} was set at 20 $\mu g/m^3$ (around 1 million cells). Those values where chosen in order to limit the simulation domain to the boundaries of the principal urban area.

Figure 3.27 shows the odds ratio resulting for the NO_2 calculation for the "high exposure"scenario. The odds ratio intervals are underestimated by LUR for both Q1, Q2 and Q3, with a difference that increases with the quartile considered.

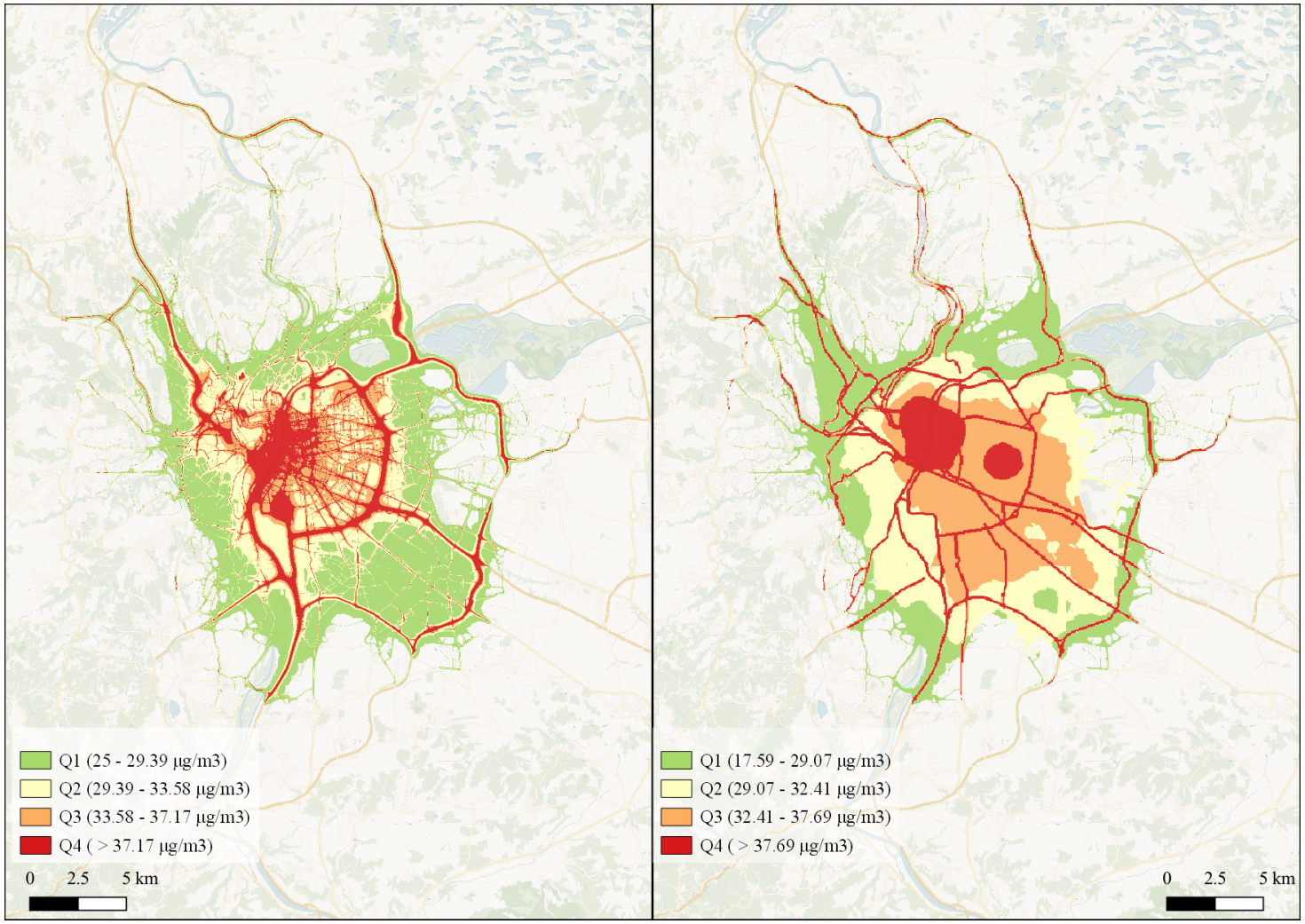

Figure 3.26: Spatial division of the domain in function of the density-weighted quartiles for NO_2, for SIRANE (on the left) and LUR (on the right) for the "high exposure"scenario

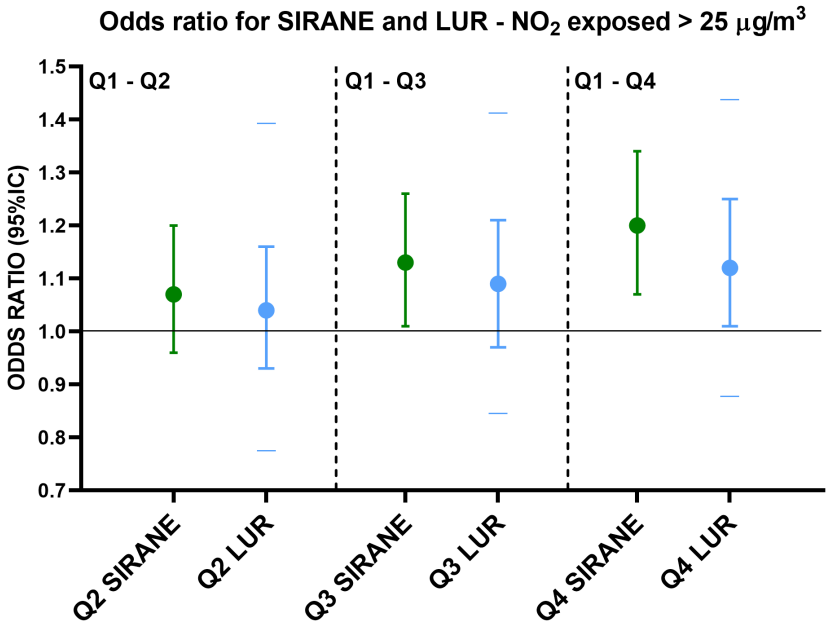

Figure 3.27: Inter-quartile odds ratios for SIRANE (fixed, in green) and for LUR (blue) for "high exposure"scenario, NO_2

	min	min (mean)	estimated	max (mean)	max
Q1	-	-	1	-	-
Q2	0.78	0.93	1.04	1.16	1.39
Q3	0.84	0.97	1.09	1.21	1.41
Q4	0.87	1.01	1.12	1.25	1.43

Table 3.29: Inter-quartile odds ratios estimated by LUR for the NO_2 in the "high exposure"scenario

The significance of the odds ratio between Q1 and Q3 is barely captured, with more than three out of four simulations indicating no significant values (figure 3.28). For the Q1-Q4, half of simulations captured the significance, with the great majority of odds ratios values under the reference (219 on 250).

Results for the PM_{10} are showed in table 3.29 and in figure 3.29:

It is observable from figure 3.29 and table 3.30 that the LUR highly un-

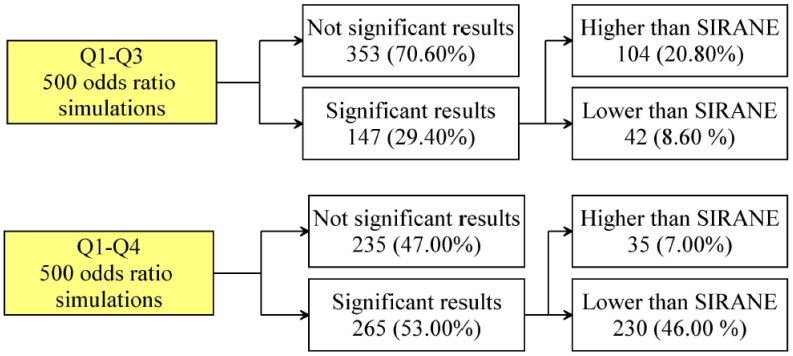

Figure 3.28: Percentages of significance lost by the LUR for the NO_2 inter-quartile odds ratios between Q1-Q3 and Q1-Q4 in the "high exposure"scenario

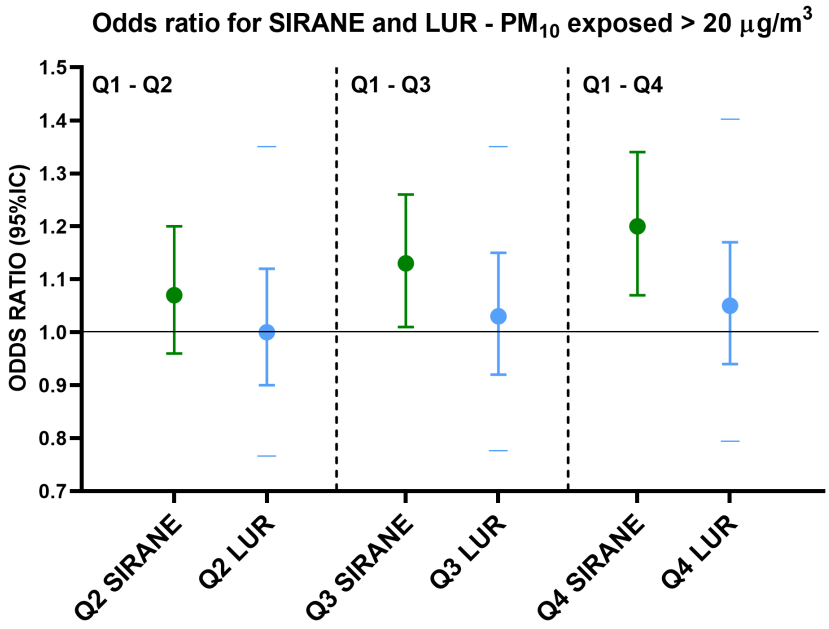

Figure 3.29: Inter-quartile odds ratios for SIRANE (reference, in green) and for LUR (blue); values for PM_{10} in the "high exposure"scenario

derestimate values for all the quartiles considered. Averaged odds ratios are barely over the unity for Q4 and Q3, while for Q2 no effect was observed. Figure 3.30 shows the loss of statistical significance for this case:

	min	min (mean)	estimated	max (mean)	max
Q1	-	-	1	-	-
Q2	0.77	0.90	1.00	1.12	1.35
Q3	0.78	0.92	1.03	1.15	1.35
Q4	0.79	0.94	1.05	1.17	1.40

Table 3.30: Inter-quartile odds ratios estimated by LUR for the PM_{10} in the "high exposure"scenario

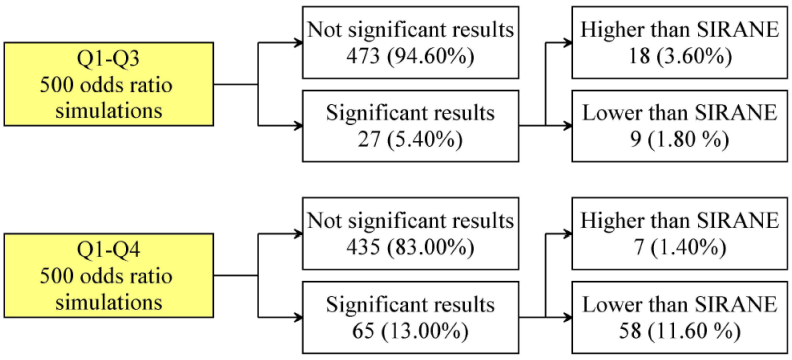

Figure 3.30: Percentages of significance lost by the LUR for the PM_{10} inter-quartile odds ratios between Q1-Q3 and Q1-Q4 in the "high exposure"scenario

As expectable by the results in table 3.30 and figure 3.29, the loss of significance is very important: only 27 values over 500 captured the significance for Q3, while 65 over 500 made it for the Q4.

The capability of the LUR model of capturing local scale pollution variability in order to calculate epidemiological outputs have been tested trough a comparison with reference values obtained applying the SIRANE model.

For the NO_2, in the base scenario the odds ratios calculated by LUR were similar to SIRANE's, with half of simulations that individuated the significance for Q1-Q3 and almost five out of six for Q1-Q4. On the other hand, in the "high exposure"one results were less promising. A similar behavior was individuated for the PM_{10} for the base scenario, while a very important loss of significant results for the "high exposure"one. This procedure is to be considered as a way of testing models' capability to capture spatial pollutant variations, both applied to great exposure ranges (for the base scenario) and smaller-ones (the "high exposure"scenario). The LUR model exhibited it can reproduce satisfactory results considering populations for which the exposure ranges are sufficiently great and inter-quartiles concentration differences are well marked. On the contrary, it demonstrates not to have a sufficient resolution to capture inter-quartile differences over the adopted thresholds at an acceptable level, especially for PM_{10}. However, the important difference for PM_{10} is also attributable to the existing bias between the two models' values within the domain.

4 Conclusions and perspectives

A relative comparison between different modelling approaches has been carried out for the city of Lyon in the years 2010 and 2000, with the objective of evaluating their interchangeability for the assignment of exposure values in retrospective nested case-control studies on cancer. The study has been developed as a part of the XENAIR project, which is one of the largest prospective studies to date investigating ambient air pollution exposure and breast cancer risk, relying on a 20 years case-control cohort named E3N located within the whole French European territory. Exposure concentrations to NO_2, O_3 and PM_{10} estimated by a street-canyon dispersion model (SIRANE) and a land use regression model (LUR) were compared for the year 2010 and 2000 over 785 subjects of the E3N cohort living in the Lyon Metropolitan Area. Visual data descriptions and statistical indicators (Pearson's r, Spearman's ρ, Cohens inter-quintile $w\kappa$) were applied, with a focus on the loss of information when passing from a deterministic model providing spatially refined estimated concentrations in a relatively small domain to a stochastic national approach. For the comparison in year 2010, also a regional CTM model (CHIMERE), a simple Nearest-Air Quality Monitoring Station (AQMS) model and three GIS-based metrics related to the proximity to roads were also involved. Specific land use types (basing on Corine Land Cover database) and socio-economic factors were considered and evaluated as possible sources for inter-model misclassifications. The impact on the estimation of a theoretical breast cancer risk of replacing the exposure values of the SIRANE model with those of the LUR model, for a virtual cohort of 10000 subjects, was assessed through an iterative procedure.

Good correlation levels were observed between SIRANE an LUR for 2010. Pearson's r were between 0.7 and 0.8 for all pollutants, Sperman's ρ were all above 0.8. Moreover, a substantial agreement was observed in term of inter-quintile reliability (Cohen's $w\kappa$ above 0.6). Lower correlation was observed with CHIMERE and the GIS-based metrics, indicating that those approaches are more suitable for preliminary analysis. The land use regression model showed a tendency in overestimating NO_2 concentrations in correspondence of a "continuous urban fabric"land use type, probably due to the high weight given to the "High density urban"predictor. Both models individuated significant decrease in NO_2 exposure for higher average income areas, defined at IRIS level, while for the PM_{10} SIRANE captured this difference more precisely. While comparing years 2000 and 2010, it was observed that he correlation and inter-quintile reliability between SIRANE and LUR were slightly higher for the year 2000. An important decrease for exposures to NO_2 and PM_{10} and an increase for those to O_3 was clearly observed in both model during this 10-year period. Nevertheless, in comparative terms the models behaved the same way. A little underestimation of inter-quartile odds ratios were observed in LUR with respect to SIRANE, both considering exposure to NO_2 and PM_{10}, over the entire Lyon Metropolitan Area. This underestimation was strongly underlined when only focusing within the urban area, especially for PM_{10}, indicating the ability to capture small-scale variation over a highly exposed population as a limitation of the land use regression model.

In summary, the stochastic approach of the land use regression model has been evaluated as a possible alternative to a deterministic dispersion models. The loss of information observed can be considered as acceptable considering the advantages provided (a greatly wider domain, a lower amount of input data requested and an inferior computational cost). However, several limitations emerged, regarding the ability of predict small-scale variations over highly exposed population and the low capacity of describing ozone spatial distribution. Further improvements are surely needed in the developing of this kind of models.

Future perspectives

This study is considered as integral part of the XENAIR project and will be the subject of a scientific publication during next months.

As explained, XENAIR is an innovative project, considering a cohort of 10000 subjects and exposure to 8 pollutants within a 20-years period. Consequently, a secondary aim of this study was to provide a standard methodology to be applied for the comparison of the two models within the whole reference period (1990-2010). One of the future developments of the project is also to consider exposures before the year 1990, so that a greater overview over historical exposure would be set, with a focus on periods of higher vulnerability. Women belonging to the E3N cohort will be interviewed with other questionnaires in order to have a complete reconstruction of their historical residences (QHR project). One of the limitations of residential geocoding is the assumption of considering as exposure concentration the value at the residential address. There is a great interest in considering a wider perspective that also includes exposures to air pollution in different locations, for example the workplace. The project APoPCo (Atmospheric POllution and Physical Activity linked with COmmute) develops by side of the project XENAIR with the objective to analyze the association between ambient air pollution exposure and breast cancer risk, within a nested case-control cohort (E3N) for 1990-2010, considering the exposition at the professional address and during the commute in addition to the residential one. Within the project APoPCo, the choosing of the commute transport type and the eventual physical activity linked to it are also taken into account (it is widely demonstrated that a regular physical activity lower the risk of breast cancer [Thune et al., 1997]).

Furthermore, analyses between the exposure values given by SIRANE and LUR will be performed also under the APoPCo project, considering residential, commute and workplace exposure. The future application of SIRANE in other cities of France could be also a great opportunity to assess the LUR performances considering different simulation domains.

Bibliography

Amadou, A., Coudon, T., Praud, D., et al. (2019). Chronic low-dose exposure to xenoestrogen ambient air pollutants and risk of breast cancer: study protocol of XENAIR Project. page 32.

Andersen, Z. J., Ravnskjær, L., Andersen, K. K., et al. (2017a). Long-term exposure to fine particulate matter and breast cancer incidence in the danish nurse cohort study. *Cancer Epidemiology Biomarkers & Prevention*, 26(3):428–430.

Andersen, Z. J., Stafoggia, M., Weinmayr, G., et al. (2017b). Long-Term Exposure to Ambient Air Pollution and Incidence of Postmenopausal Breast Cancer in 15 European Cohorts within the ESCAPE Project. *Environmental Health Perspectives*, 125(10):107005.

Beelen, R., Hoek, G., Vienneau, D., et al. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmospheric Environment*, 72:10–23.

Beelen, R., Voogt, M., Duyzer, J., Zandveld, P., and Hoek, G. (2010). Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. *Atmospheric Environment*, 44(36):4614–4621.

Benson, P. (1988). Development and Verification of the California Line Source Dispersion Model. *Transportation research Records*, 1176.

Bernstein, J. A., Alexis, N., Barnes, C., et al. (2004). Health effects of air pollution. *Journal of Allergy and Clinical Immunology*, 114(5):1116–1123.

Binachon, B., Dossus, L., Danjou, A. M. N., Clavel-Chapelon, F., and Fervers, B. (2014). Life in urban areas and breast cancer risk in the French E3N cohort. *European Journal of Epidemiology*, 29(10):743–751.

Bolte, G., Tamburlini, G., and Kohlhuber, M. (2010). Environmental inequalities among children in europe–evaluation of scientific evidence and policy implications. *The European Journal of Public Health*, 20(1):14–20.

Bonner, M. R., Han, D., Nie, J., et al. (2003). Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology*, 14(4):408–412.

Bray, F., Ferlay, J., Soerjomataram, I., et al. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Briggs, D. J., Collins, S., Elliott, P., et al. (1997). Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science*, 11(7):699–718.

Briggs, D. J., de Hoogh, C., Gulliver, J., et al. (2000). A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of The Total Environment*, 253(1-3):151–167.

Brody, J. G., Moysich, K. B., Humblet, O., et al. (2007). Environmental pollutants and breast cancer: Epidemiologic studies. *Cancer*, 109(S12):2667–2711.

Büttner, G. (2014). CORINE Land Cover and Land Cover Change Products. In Manakos, I. and Braun, M., editors, *Land Use and Land Cover*

Mapping in Europe, volume 18, pages 55–74. Springer Netherlands, Dordrecht. Series Title: Remote Sensing and Digital Image Processing.

Chang, J. C. and Hanna, S. R. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, 87(1-3).

Cimorelli, A. J., Perry, S. G., Venkatram, A., et al. (2005). AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model Formulation and Boundary Layer Characterization. *Journal of Applied Meteorology*, 44(5):682–693.

Clavel-Chapelon, F. and E3N Study Group (2015). Cohort Profile: The French E3N Cohort Study. *International Journal of Epidemiology*, 44(3):801–809.

Colette, A., Granier, C., Hodnebrog, ., et al. (2011). Air quality trends in europe over the past decade: a first multi-model assessment. *Atmospheric Chemistry and Physics*, 11(22):11657–11678.

Colvile, R., Hutchinson, E., Mindell, J., and Warren, R. (2001). The transport sector as a source of air pollution. *Atmospheric Environment*, 35(9):1537–1565.

Coogan, P. F., White, L. F., Yu, J., et al. (2016). Long term exposure to NO 2 and diabetes incidence in the Black Women's Health Study. *Environmental Research*, 148:360–366.

Coudon, T., Danjou, A. M. N., Faure, E., et al. (2019a). Development and performance evaluation of a GIS-based metric to assess exposure to airborne pollutant emissions from industrial sources. *Environmental Health*, 18(1):8.

Coudon, T., Grassot, L., Dubuis, M., et al. (2019b). Assessment of long-term exposure to airborne pollution in france (1990-2010). *Environmental Epidemiology*.

Cyrys, J., Hochadel, M., Gehring, U., et al. (2005). GIS-Based Estimation of Exposure to Particulate Matter and NO2 in an Urban Area: Stochastic versus Dispersion Modeling. *Environmental Health Perspectives*, 113(8):987–992.

Dadvand, P., Ostro, B., Figueras, F., et al. (2014). Residential Proximity to Major Roads and Term Low Birth Weight: The Roles of Air Pollution, Heat, Noise, and Road-Adjacent Trees. *Epidemiology*, 25(4):518–525.

Danjou, A. M. N., Coudon, T., Praud, D., et al. (2019). Long-term airborne dioxin exposure and breast cancer risk in a case-control study nested within the French E3N prospective cohort. *Environment International*, 124:236–248.

Danjou, A. M. N., Fervers, B., Boutron-Ruault, M.-C., et al. (2015). Estimated dietary dioxin exposure and breast cancer risk among women from the French E3N prospective cohort. *Breast cancer research : BCR*, 17:39.

Dėdelė, A. and Miškinytė, A. (2015). The statistical evaluation and comparison of ADMS-Urban model for the prediction of nitrogen dioxide with air quality monitoring network. *Environmental Monitoring and Assessment*, 187(9):578.

de Hoogh, K., Korek, M., Vienneau, D., et al. (2014). Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. *Environment International*, 73:382–392.

Deguen, S. and Zmirou-Navier, D. (2010). Social inequalities resulting from health risks related to ambient air quality–a european review. *The European Journal of Public Health*, 20(1):27–35.

Doherty, R. M., Stevenson, D. S., Collins, W. J., and Sanderson, M. G. (2005). Influence of convective transport on tropospheric ozone and its precursors in a chemistry-climate model. *Atmos. Chem. Phys.*, page 14.

EPA (2005). Revision to the Guideline on Air Quality Models: Adoption of a Preferred General Purpose (Flat and Complex Terrain) Dispersion Model and Other Revisions. 70(216):68218–68261.

Ezzati, M. and Organització Mundial de la Salut (2004). *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*. World Health Organization, Geneva. OCLC: 803774407.

Fairburn, J., Schüle, S. A., Dreger, S., Karla Hilz, L., and Bolte, G. (2019). Social Inequalities in Exposure to Ambient Air Pollution: A Systematic Review in the WHO European Region. *International Journal of Environmental Research and Public Health*, 16(17):3127.

Faure, E., Danjou, A. M., Clavel-Chapelon, F., et al. (2017). Accuracy of two geocoding methods for geographic information system-based exposure assessment in epidemiological studies. *Environmental Health*, 16.

Forastiere, F., Renzi, M., Cesaroni, G., and Stafoggia, M. (2019). Low-level air pollution and natural cause mortality in Rome: comparison of results based on European wide and local land use regression models within the ELAPSE project. *Environmental Epidemiology*, 3:125.

Forouzanfar, M. H., Afshin, A., Alexander, L. T., et al. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1659–1724.

Fournier, A., Berrino, F., and Clavel-Chapelon, F. (2008). Unequal risks for breast cancer associated with different hormone replacement therapies: results from the E3N cohort study. *Breast Cancer Research and Treatment*, 107(1):103–111.

Gilbert, M. and Wendell, E. (2014). *Introduction to Environmental Engineering and Science*. Pearson, third edition edition.

Gray, S. C., Edwards, S. E., and Miranda, M. L. (2010). Assessing exposure metrics for PM and birth weight models. *Journal of Exposure Science & Environmental Epidemiology*, 20(5):469–477.

Guerreiro, C. B., Foltescu, V., and de Leeuw, F. (2014). Air quality status and trends in europe. *Atmospheric Environment*, 98:376–384.

Harvie, M., Howell, A., and Evans, D. G. (2015). Can Diet and Lifestyle Prevent Breast Cancer: What Is the Evidence? *American Society of Clinical Oncology Educational Book*, (35):e66–e73.

Hennig, F., Sugiri, D., Tzivian, L., et al. (2016). Comparison of Land-Use Regression Modeling with Dispersion and Chemistry Transport Modeling to Assign Air Pollution Concentrations within the Ruhr Area. *Atmosphere*, 7(3):48.

Henning, F., Lucht, S., and Hoffmann, B. (2018). Improvement of lur model predictions using chemistry-transport model estimates. *Environmental Health Perspectives*.

Hochadel, M., Heinrich, J., Gehring, U., et al. (2006). Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmospheric Environment*, 40(3):542–553.

Hodzic, A., Jimenez, J. L., Madronich, S., et al. (2009). Modeling organic aerosols during MILAGRO: importance of biogenic secondary organic aerosols. *Atmospheric Chemistry and Physics*, 9(18):6949–6981. Publisher: Copernicus GmbH.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., and van den Brandt, P. A. (2002). Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *The Lancet*, 360(9341):1203–1209.

Holmes, N. and Morawska, L. (2006). A review of dispersion modelling and its application to the dispersion of particles: An overview of differ-

ent dispersion models available. *Atmospheric Environment*, 40(30):5902–5928.

IGN (2019). Bd topo® version 3.0 - descriptif de contenu. pages 319–320.

Jemal, A., Center, M. M., DeSantis, C., and Ward, E. M. (2010). Global Patterns of Cancer Incidence and Mortality Rates and Trends. *Cancer Epidemiology Biomarkers & Prevention*, 19(8):1893–1907.

Jerrett, M., Burnett, R. T., Pope, C. A., et al. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, 360(11):1085–1095.

Johnson, M., Isakov, V., Touma, J., Mukerjee, S., and Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, 44(30):3660–3668.

Kampa, M. and Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367.

Künzli, N., Jerrett, M., Mack, W. J., et al. (2005). Ambient Air Pollution and Atherosclerosis in Los Angeles. *Environmental Health Perspectives*, 113(2):201–206.

Kottegoda, N. T. and Rosso, R. (2008). *Applied statistics for civil and environmental engineers*. Blackwell Pub, second edition.

Landrigan, P. J., Fuller, R., Acosta, N. J. R., et al. (2018). The Lancet Commission on pollution and health. *The Lancet*, 391(10119):462–512.

Landrigan, P. J., Fuller, R., Fisher, S., et al. (2019). Pollution and children's health. *Science of The Total Environment*, 650:2389–2394.

Lebret, E., Briggs, D., van Reeuwijk, H., et al. (2000). Small area variations in ambient NO2 concentrations in four European areas. *Atmospheric Environment*, 34(2):177–185.

Lelieveld, J., Pozzer, A., Pöschl, U., et al. (2020). Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular Research*.

Liu, Z., Guan, Q., Luo, H., et al. (2019). Development of land use regression model and health risk assessment for NO2 in different functional areas: A case study of Xi'an, China. *Atmospheric Environment*, 213:515–525.

Loomis, D., Grosse, Y., Lauby-Secretan, B., et al. (2014). The carcinogenicity of outdoor air pollution. *Lancet Oncology*, 14(13):1262–1263.

Maheswaran, R. and Elliott, P. (2003). Stroke Mortality Associated With Living Near Main Roads in England and Wales: A Geographical Study. *Stroke*, 34(12):2776–2780.

Marshall, J. D., Nethery, E., and Brauer, M. (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, 42(6):1359–1369.

Meleux, F., Solmon, F., and Giorgi, F. (2007). Increase in summer european ozone amounts due to climate change. *Atmospheric Environment*, 41(35):7577–7587.

Menut, L., Bessagnet, B., Khvorostyanov, D., et al. (2013). CHIMERE 2013: a model for regional atmospheric composition modelling. *Geosci. Model Dev.*, 6(4):981–1028.

Miyake, Y., Yura, A., and Iki, M. (2002). Relationship between Distance from Major Roads and Adolescent Health in Japan. *Journal of Epidemiology*, 12(6):418–423.

Monks, P. S., Archibald, A. T., Colette, A., et al. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric Chemistry and Physics*, 15(15):8889–8973.

Morelli, X., Gabet, S., Rieux, C., et al. (2019). Which decreases in air pollution should be targeted to bring health and economic benefits and improve environmental justice? *Environment International*, 129:538–550.

Nerriere, l., Zmirou-Navier, D., Blanchard, O., et al. (2005). Can we use fixed ambient air monitors to estimate population long-term exposure to air pollutants? The case of spatial variability in the Genotox ER study. *Environmental Research*, 97(1):32–42.

Nie, J., Beyea, J., Bonner, M. R., et al. (2007). Exposure to traffic emissions throughout life and risk of breast cancer: the Western New York Exposures and Breast Cancer (WEB) study. *Cancer Causes & Control*, 18(9):947–955.

Nuckols John R., Ward Mary H., and Jarup Lars (2004). Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. *Environmental Health Perspectives*, 112(9):1007–1015. Publisher: Environmental Health Perspectives.

Pahlow, M., Parlange, M. B., and Porté-Agel, F. (2001). On Monin–Obukhov Similarity In The Stable Atmospheric Boundary Layer. *Boundary-Layer Meteorology*, 99(2):225–248.

Pless-Mulloli, T., Phillimore, P., and Moffatt, S. (1998). Lung cancer, proximity to industry, and poverty in northeast England. *Environmental Health Perspectives*, 106(4):8.

Potischman, N. and Troisi, R. (1999). In-utero and Early Life Exposures in Relation to Risk of Breast Cancer. *Cancer Causes & Control*, 10(6):561–573.

Queen, A. and Zhang, Y. (2008). Examining the sensitivity of MM5–CMAQ predictions to explicit microphysics schemes and horizontal grid resolutions, Part III—The impact of horizontal grid resolution. *Atmospheric Environment*, 42(16):3869–3881.

Raaschou-Nielsen, O., Beelen, R., Wang, M., et al. (2016). Particulate matter air pollution components and risk for lung cancer. *Environment International*, 87:66–73.

Ragettli, M., Tsai, M.-Y., Braun-Fahrländer, C., et al. (2014). Simulation of Population-Based Commuter Exposure to NO2 Using Different Air Pollution Models. *International Journal of Environmental Research and Public Health*, 11(5):5049–5068.

Ryan, P. H. and LeMasters, G. K. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology*, 19(sup1):127–133.

Salizzoni, P., Soulhac, L., and Mejean, P. (2009). Street canyon ventilation and atmospheric turbulence. *Atmospheric Environment*, 43(32):5056–5067.

Schaap, M., Cuvelier, C., Hendriks, C., et al. (2015). Performance of European chemistry transport models as function of horizontal resolution. *Atmospheric Environment*, 112:90–105.

Schaap, M., Vautard, R., Bergström, R., et al. (2007). Evaluation of long-term aerosol simulations from seven air quality models and their ensemble in the EURODELTA study. *Atmos. Environ*, 41:2083–2097.

Sellier, Y., Galineau, J., Hulin, A., et al. (2014). Health effects of ambient air pollution: Do different methods for estimating exposure lead to different results? *Environment International*, 66:165–173.

Sicard, P., Serra, R., and Rossello, P. (2016). Spatiotemporal trends in ground-level ozone concentrations and metrics in france over the time period 1999–2012. *Environmental Research*, 149:122–144.

Soulhac, L., Garbero, V., Salizzoni, P., Mejean, P., and Perkins, R. (2009). Flow and dispersion in street intersections. *Atmospheric Environment*, 43(18):2981–2996.

Soulhac, L., Nguyen, C. V., Volta, P., and Salizzoni, P. (2017). The model SIRANE for atmospheric urban pollutant dispersion. PART III: Validation against NO2 yearly concentration measurements in a large urban agglomeration. *Atmospheric Environment*, 167:377–388.

Soulhac, L., Salizzoni, P., Cierco, F.-X., and Perkins, R. (2011). The model SIRANE for atmospheric urban pollutant dispersion; part I, presentation of the model. *Atmospheric Environment*, 45(39):7379–7395.

Soulhac, L., Salizzoni, P., Mejean, P., Didier, D., and Rios, I. (2012). The model SIRANE for atmospheric urban pollutant dispersion; PART II, validation of the model on a real case study. *Atmospheric Environment*, 49:320–337.

Tchepel, O., Costa, A., Martins, H., et al. (2010). Determination of background concentrations for air quality models using spectral analysis and filtering of monitoring data. *Atmospheric Environment*, 44(1):106–114.

Thune, I., Brenn, T., Lund, E., and Gaard, M. (1997). Physical activity and the risk of breast cancer. *The New England Journal of Medicine*, page 7.

Vardoulakis, S., Fisher, B. E., Pericleous, K., and Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: a review. *Atmospheric Environment*, 37(2):155–182.

Vautard, R., Beekmann, M., Roux, J., and Gombert, D. (2001). Validation of a hybrid forecasting system for the ozone concentrations over the Paris area. *Atmospheric Environment*, 35(14):2449–2461.

Venn, A., Yemaneberhan, H., Lewis, S., Parry, E., and Britton, J. (2005). Proximity of the Home to Roads and the Risk of Wheeze in an Ethiopian Population. *Occupational and Environmental Medicine*, 62(6):376–380.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

Wang, M., Beelen, R., Eeftens, M., et al. (2012). Systematic Evaluation of Land Use Regression Models for NO2. *Environmental Science & Technology*, 46(8):4481–4489.

Wang Meng, Gehring Ulrike, Hoek Gerard, et al. (2015). Air Pollution and Lung Function in Dutch Children: A Comparison of Exposure Estimates and Associations Based on Land Use Regression and Dispersion Exposure Modeling Approaches. *Environmental Health Perspectives*, 123(8):847–851.

Ward, M. H., Nuckols, J. R., Giglierano, J., et al. (2005). Positional Accuracy of Two Methods of Geocoding:. *Epidemiology*, 16(4):542–547.

Weinmayr, G., Pedersen, M., Stafoggia, M., et al. (2018). Particulate matter air pollution components and incidence of cancers of the stomach and the upper aerodigestive tract in the european study of cohorts of air pollution effects (escape). *Environment International*, 120:163–171.

White, A. J., Bradshaw, P. T., and Hamra, G. B. (2018). Air pollution and breast cancer: a review. *Current Epidemiology Reports*, 5(2):92–100.

WHO (2005). Who air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide.

WHO (2016). *Ambient air pollution: A global assessment of exposure and burden fo disease*.

WHO (2017). Don't pollute my future! The impact of the environment on children's health.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19:251–253.

Xie, Y., Zhao, B., Zhang, L., and Luo, R. (2015). Spatiotemporal variations of pm2.5 and pm10 concentrations between 31 chinese cities and their relationships with so2, no2, co and o3. *Particuology*, 20:141–149.

Yaghjyan, L., Arao, R., Brokamp, C., et al. (2017). Association between air pollution and mammographic breast density in the Breast Cancer Surveilance Consortium. *Breast Cancer Research*, 19(1):36.

Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*, 7(1):37.

Zou, B., Wilson, J. G., Zhan, F. B., and Zeng, Y. (2009). Air pollution exposure assessment methods utilized in epidemiological studies. *Journal of Environmental Monitoring*, 11(3):475–490. Publisher: The Royal Society of Chemistry.

Zyryanov, D., Foret, G., Eremenko, M., et al. (2012). 3-D evaluation of tropospheric ozone simulations by an ensemble of regional Chemistry Transport Model. *Atmospheric Chemistry and Physics*, 12(7):3219–3240.

Appendices

Appendix A: Virtual population plots

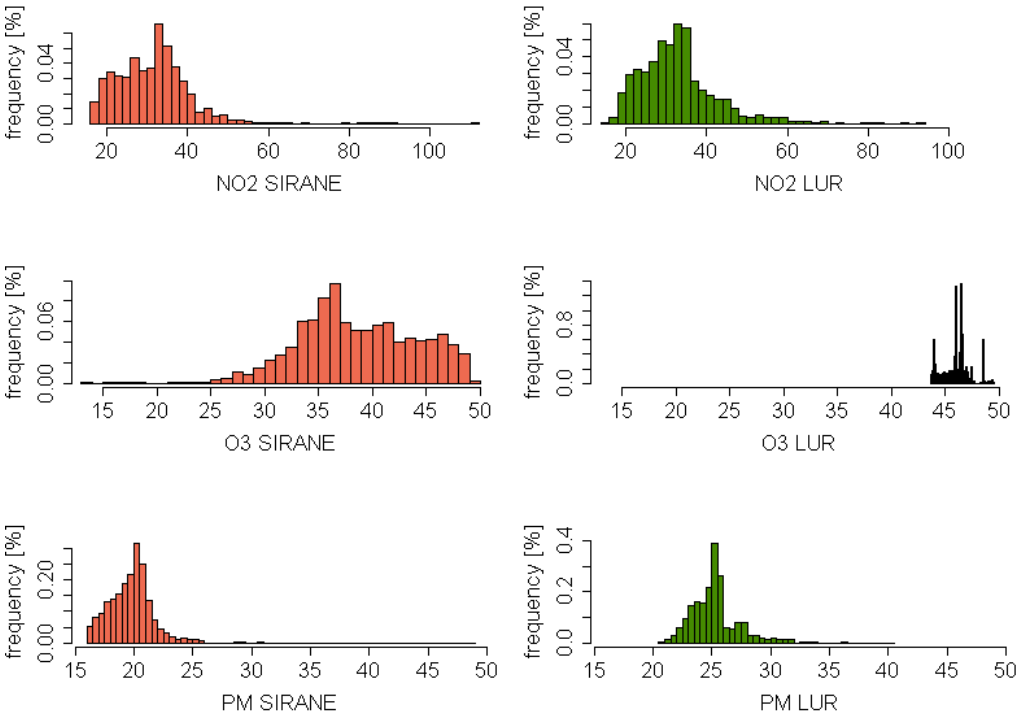

Figure A.1: Histograms - SR1 population

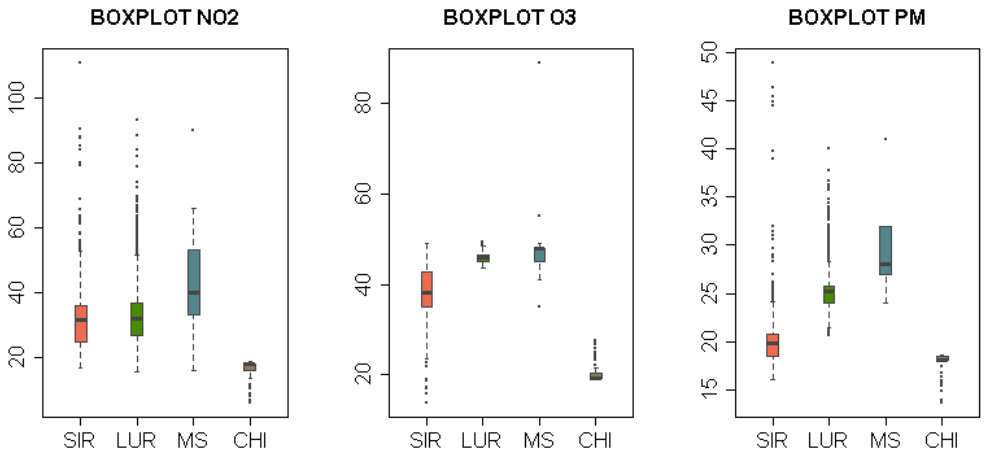

Figure A.2: Boxplots - SR1 population

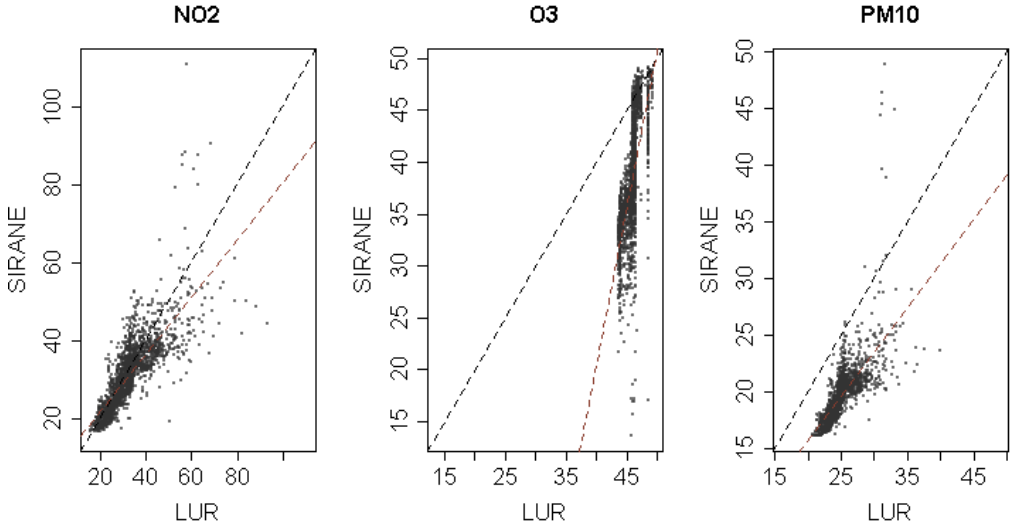

Figure A.3: Scatterplots - SR1 population

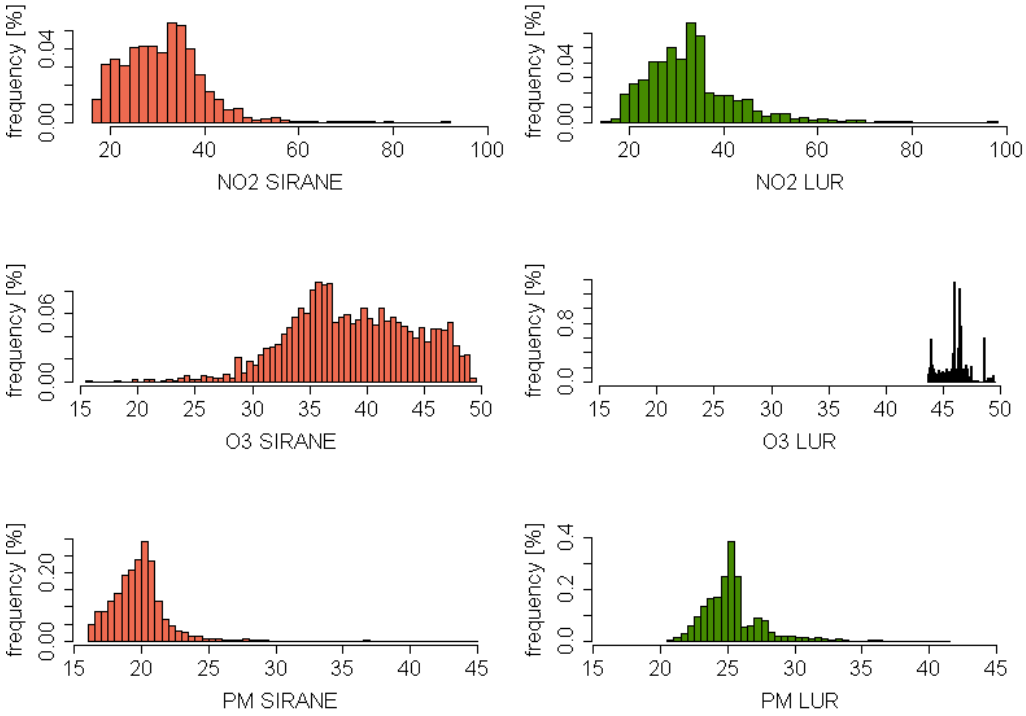

Figure A.4: Histograms - SR2 population

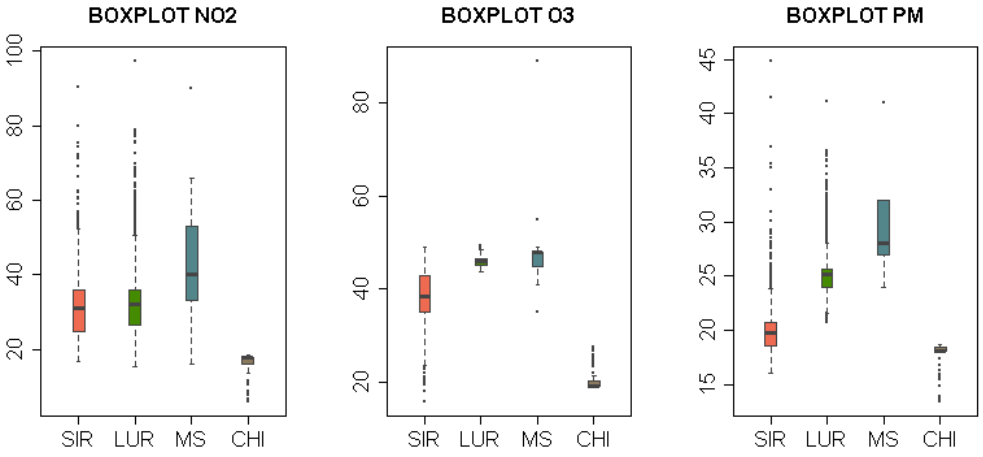

Figure A.5: Boxplots - SR2 population

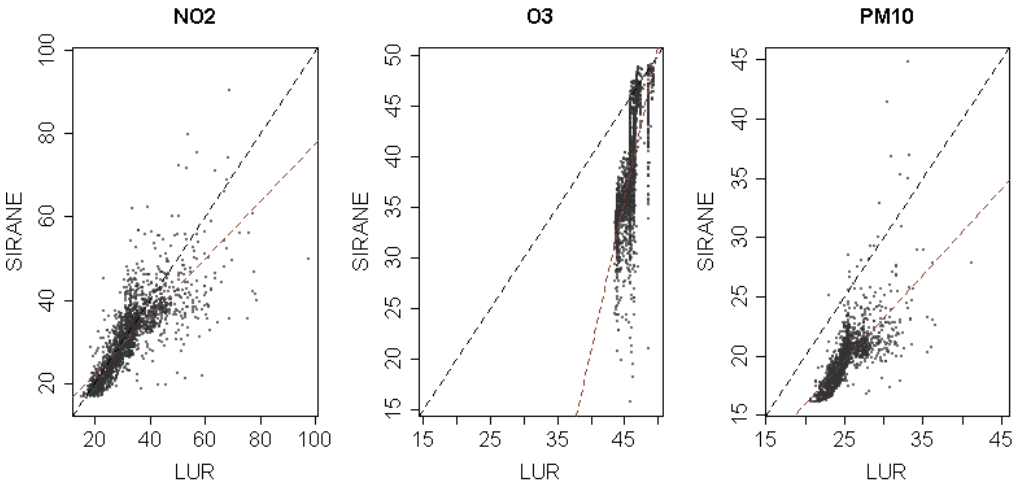

Figure A.6: Scatterplots - SR2 population

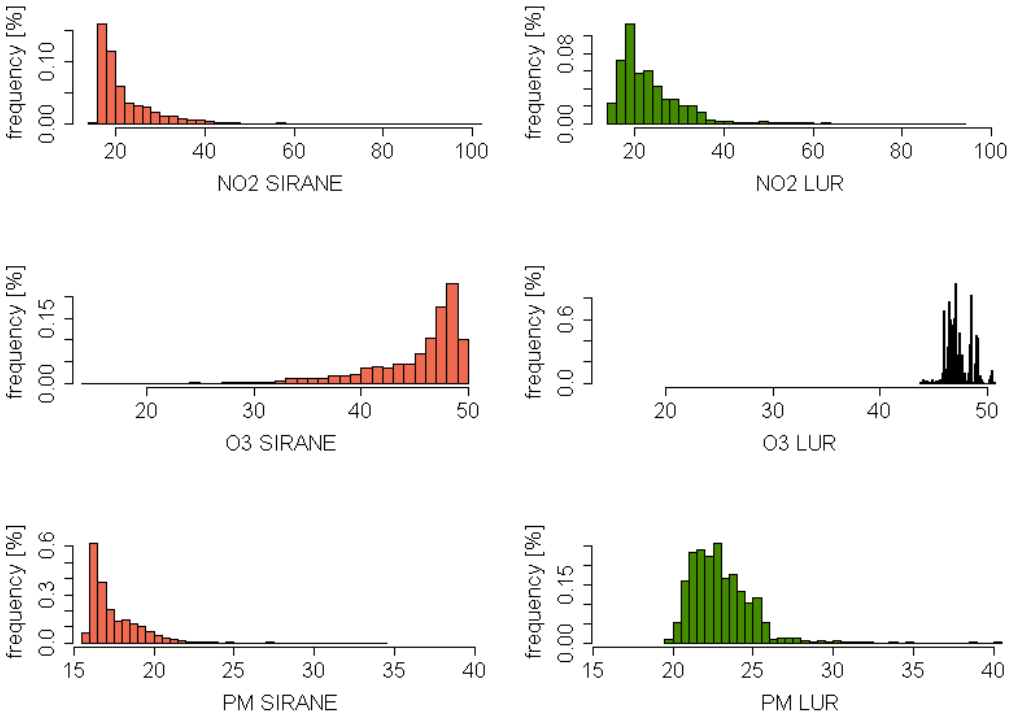

Figure A.7: Histograms - RANDOM population

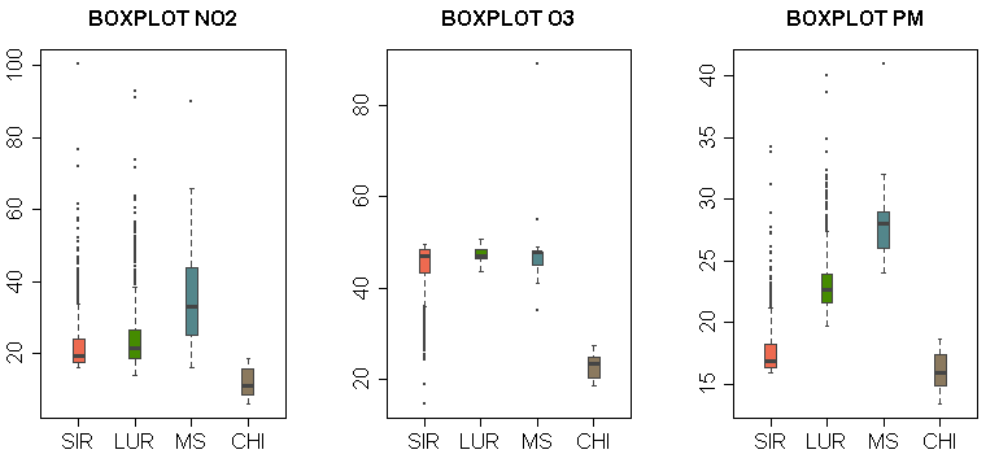

Figure A.8: Boxplots - RANDOM population

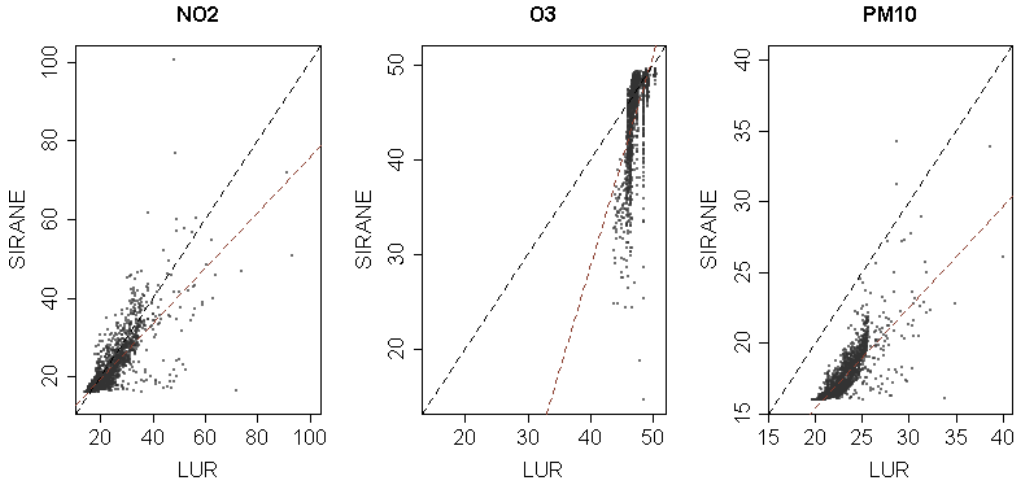

Figure A.9: Scatterplots - RANDOM population

Appendix B: Agreement coefficients - virtual populations

PEARSON's r	NO2	O3	PM	SPEARMAN's ρ	NO2	O3	PM
SIRANE-LUR	0,79	0,70	0,71	SIRANE-LUR	0,86	0,79	0,84
SIRANE-AQMS	0,05	-0,13	-0,06	SIRANE-AQMS	0,20	-0,14	-0,05
SIRANE-CHIMERE	0,53	0,58	0,42	SIRANE-CHIMERE	0,52	0,46	0,47
LUR-AQMS	0,08	-0,25	-0,04	LUR-AQMS	0,20	-0,16	-0,04
LUR-CHIMERE	0,52	0,50	0,48	LUR-CHIMERE	0,52	0,48	0,51
CHIMERE-AQMS	0,15	0,11	-0,03	CHIMERE-AQMS	0,33	0,18	0,27

Table B.1: Correlation coefficients for SR1

PEARSON's r	NO2	O3	PM	SPEARMAN's ρ	NO2	O3	PM
SIRANE-LUR	0,78	0,70	0,71	SIRANE-LUR	0,85	0,78	0,84
SIRANE-AQMS	0,03	-0,15	-0,08	SIRANE-AQMS	0,18	-0,15	-0,07
SIRANE-CHIMERE	0,53	0,57	0,44	SIRANE-CHIMERE	0,52	0,46	0,49
LUR-AQMS	0,08	-0,24	-0,04	LUR-AQMS	0,19	-0,16	-0,04
LUR-CHIMERE	0,51	0,50	0,46	LUR-CHIMERE	0,52	0,49	0,52
CHIMERE-AQMS	0,18	0,09	-0,02	CHIMERE-AQMS	0,36	0,16	0,28

Table B.2: Correlation coefficient for SR2

PEARSON's r	NO2	O3	PM	SPEARMAN's ρ	NO2	O3	PM
SIRANE-LUR	0,75	0,71	0,70	SIRANE-LUR	0,84	0,79	0,85
SIRANE-AQMS	0,11	-0,13	-0,01	SIRANE-AQMS	0,23	-0,07	0,03
SIRANE-CHIMERE	0,58	0,61	0,48	SIRANE-CHIMERE	0,65	0,60	0,61
LUR-AQMS	0,06	-0,29	0,02	LUR-AQMS	0,14	-0,08	0,04
LUR-CHIMERE	0,53	0,49	0,51	LUR-CHIMERE	0,59	0,56	0,63
CHIMERE-AQMS	0,20	0,11	0,00	CHIMERE-AQMS	0,36	0,16	0,21

Table B.3: Correlation coefficients for RND

Cohen's wk	NO2	95%CI	O3	95%CI	PM10	95%CI
SIRANE-LUR	0,68	0,66 - 0,7	0,61	0,58 - 0,63	0,66	0,64 - 0,68
SIRANE-STAT	0,14	0,11 - 0,17	-0,12	-0,15 - -0,1	-0,06	-0,09 - -0,04
SIRANE-CHIMERE	0,34	0,31 - 0,36	0,18	0,15 - 0,21	0,33	0,3 - 0,36
LUR-PROXY	0,20	0,16 - 0,23	-0,11	-0,14 - -0,09	-0,04	-0,07 - -0,02
LUR-CHIMERE	0,33	0,3 - 0,35	0,18	0,15 - 0,21	0,34	0,31 - 0,36
CHIMERE-PROXY	0,29	0,26 - 0,32	0,22	0,19 - 0,25	0,21	0,18 - 0,23

Table B.4: Cohen's $w\kappa$ for SR1

Cohen's wk	NO2	95%CI	O3	95%CI	PM10	95%CI
SIRANE-LUR	0,68	0,66 - 0,7	0,61	0,59 - 0,63	0,66	0,65 - 0,69
SIRANE-STAT	0,14	0,09 - 0,15	-0,12	-0,16 - -0,1	-0,06	-0,1 - -0,05
SIRANE-CHIMERE	0,34	0,31 - 0,36	0,18	0,17 - 0,23	0,33	0,32 - 0,37
LUR-PROXY	0,20	0,15 - 0,22	-0,11	-0,13 - -0,08	-0,04	-0,07 - -0,02
LUR-CHIMERE	0,33	0,3 - 0,35	0,18	0,16 - 0,22	0,34	0,33 - 0,38
CHIMERE-PROXY	0,29	0,27 - 0,33	0,22	0,16 - 0,22	0,21	0,18 - 0,23

Table B.5: Cohen's $w\kappa$ for SR2

Cohen's wk	NO2	95%CI	O3	95%CI	PM10	95%CI
SIRANE-LUR	0,68	0,24 - 0,29	0,61	-0,15 - -0,1	0,66	0,22 - 0,27
SIRANE-STAT	0,14	0,51 - 0,55	-0,12	0,49 - 0,53	-0,06	0,36 - 0,41
SIRANE-CHIMERE	0,34	0,16 - 0,21	0,18	-0,13 - -0,08	0,33	0,17 - 0,23
LUR-PROXY	0,20	0,49 - 0,53	-0,11	0,33 - 0,38	-0,04	0,42 - 0,47
LUR-CHIMERE	0,33	0,2 - 0,26	0,18	-0,02 - 0,03	0,34	0,1 - 0,17
CHIMERE-PROXY	0,29	0,2 - 0,26	0,22	-0,02 - 0,03	0,21	0,1 - 0,17

Table B.6: Cohen's $w\kappa$ for RND

Appendix C: Model results not in main text

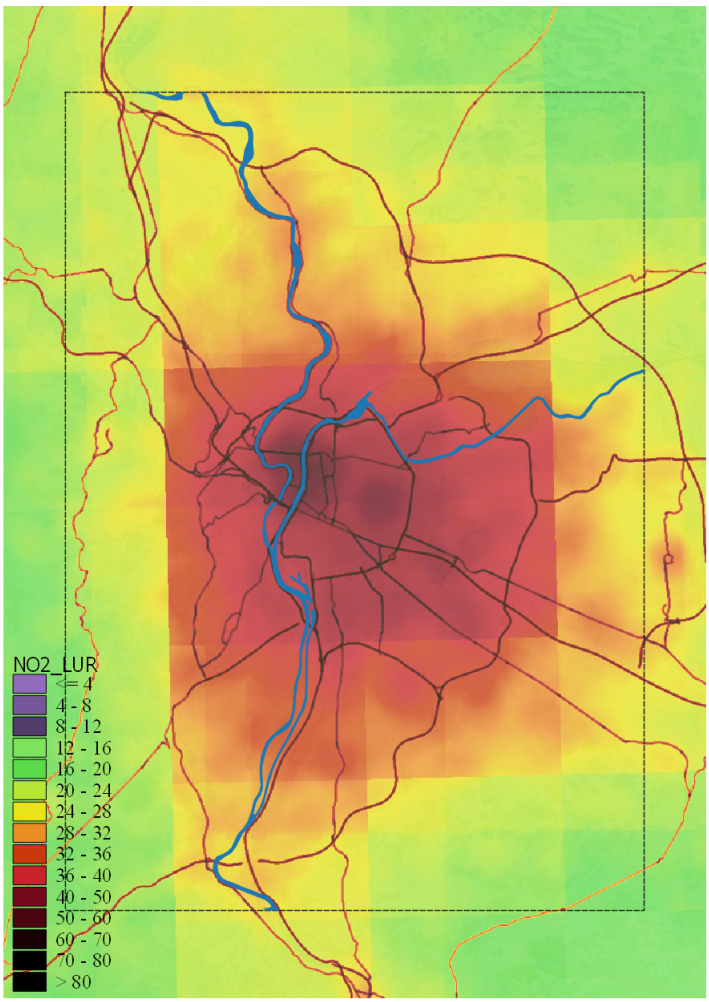

Figure C.1: LUR - NO2 - 2000

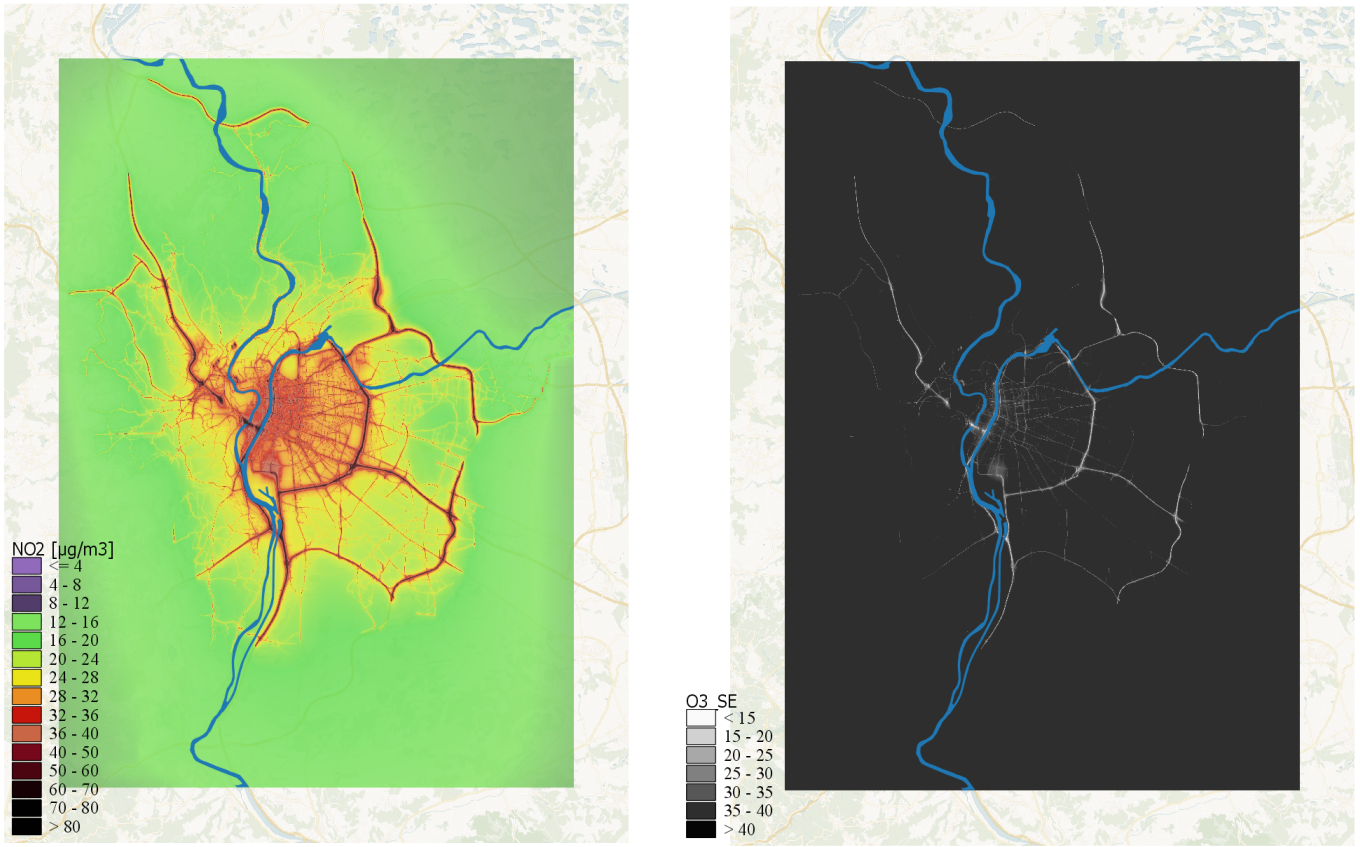

Figure C.2: SIRANE SE 2010: NO2 (right) and O3 (left)

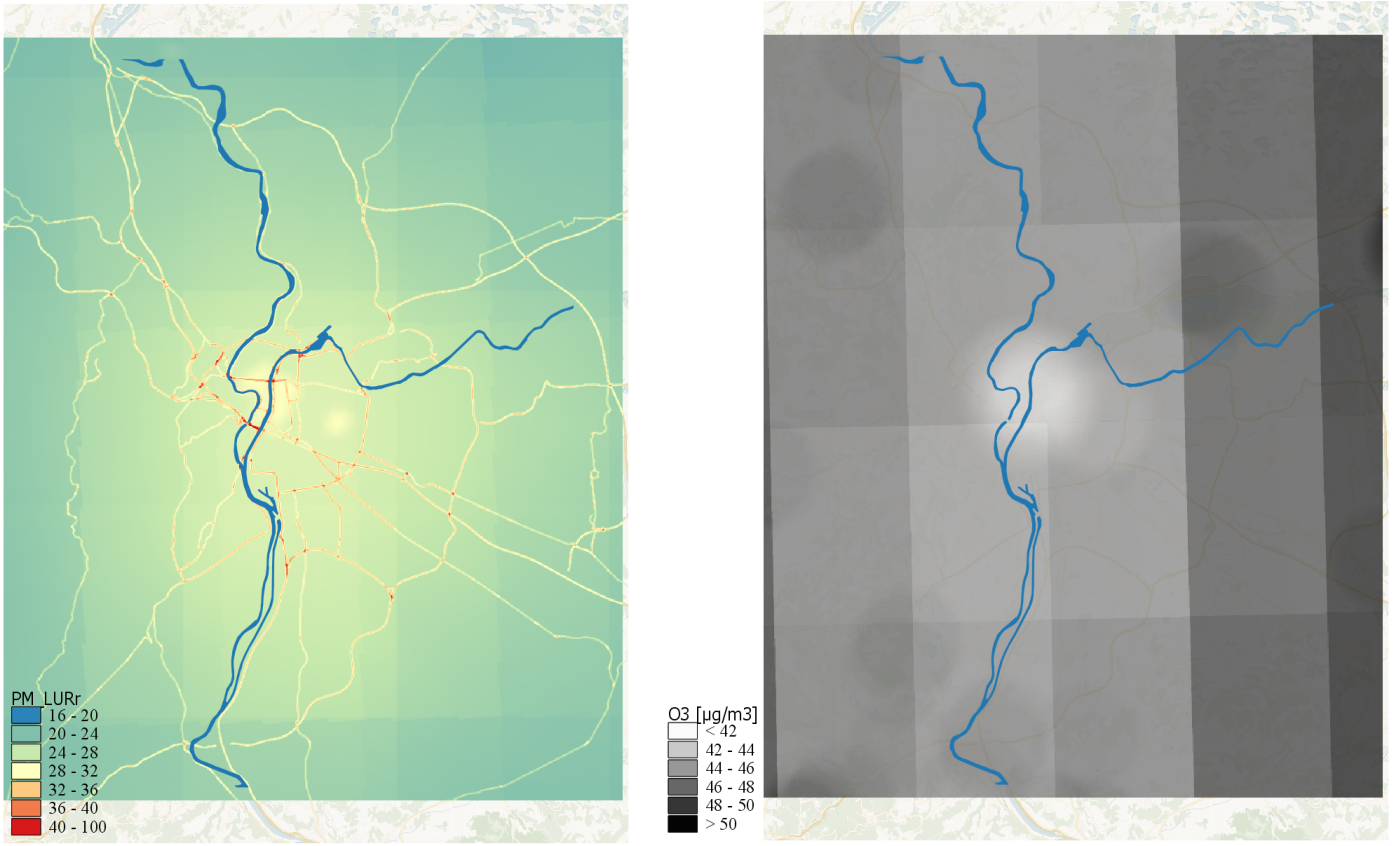

Figure C.3: LUR 2010: PM10 (right) and O3 (left)

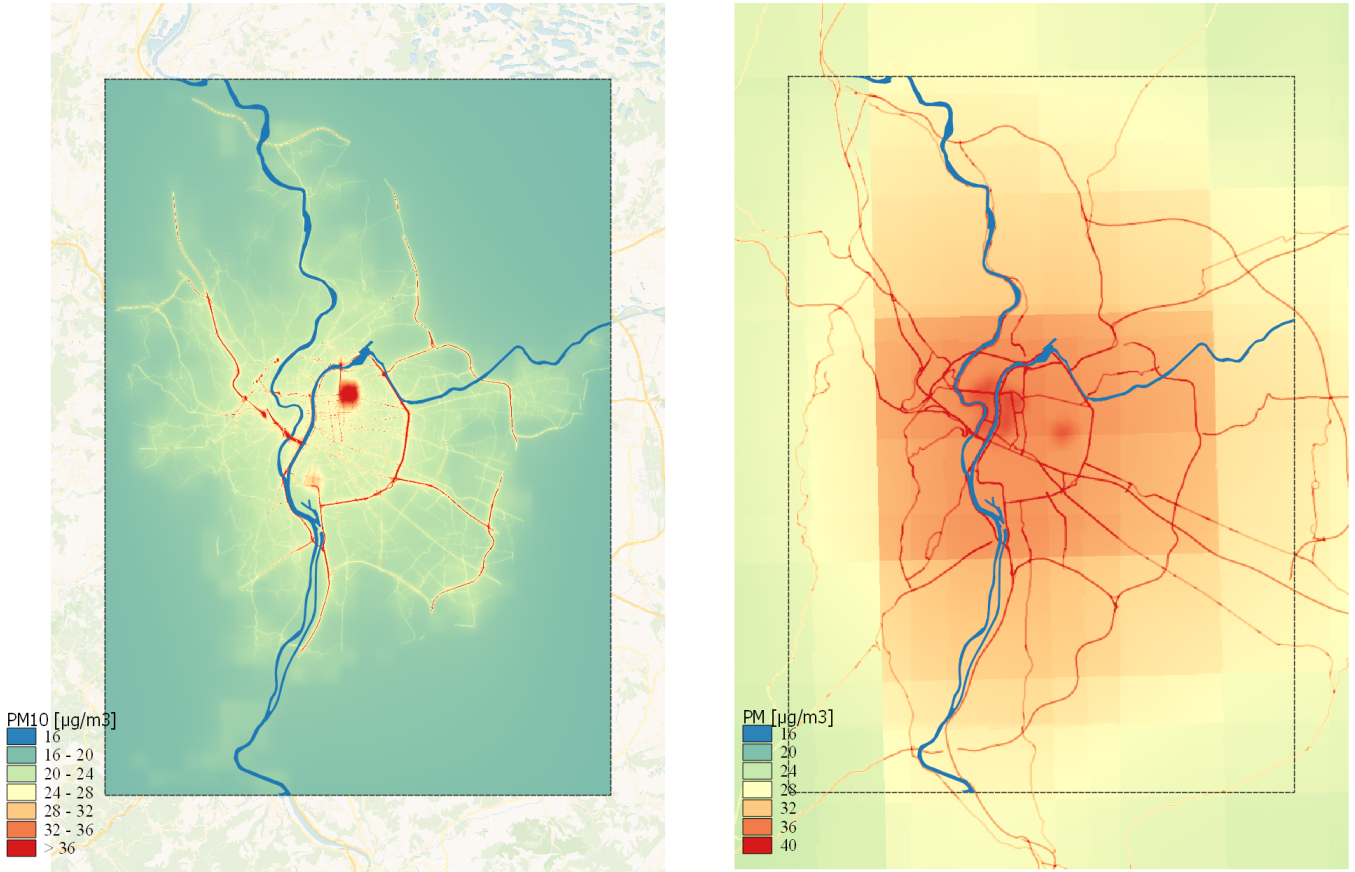

Figure C.4: PM10 - 2000: SIRANE (left) and LUR (right)

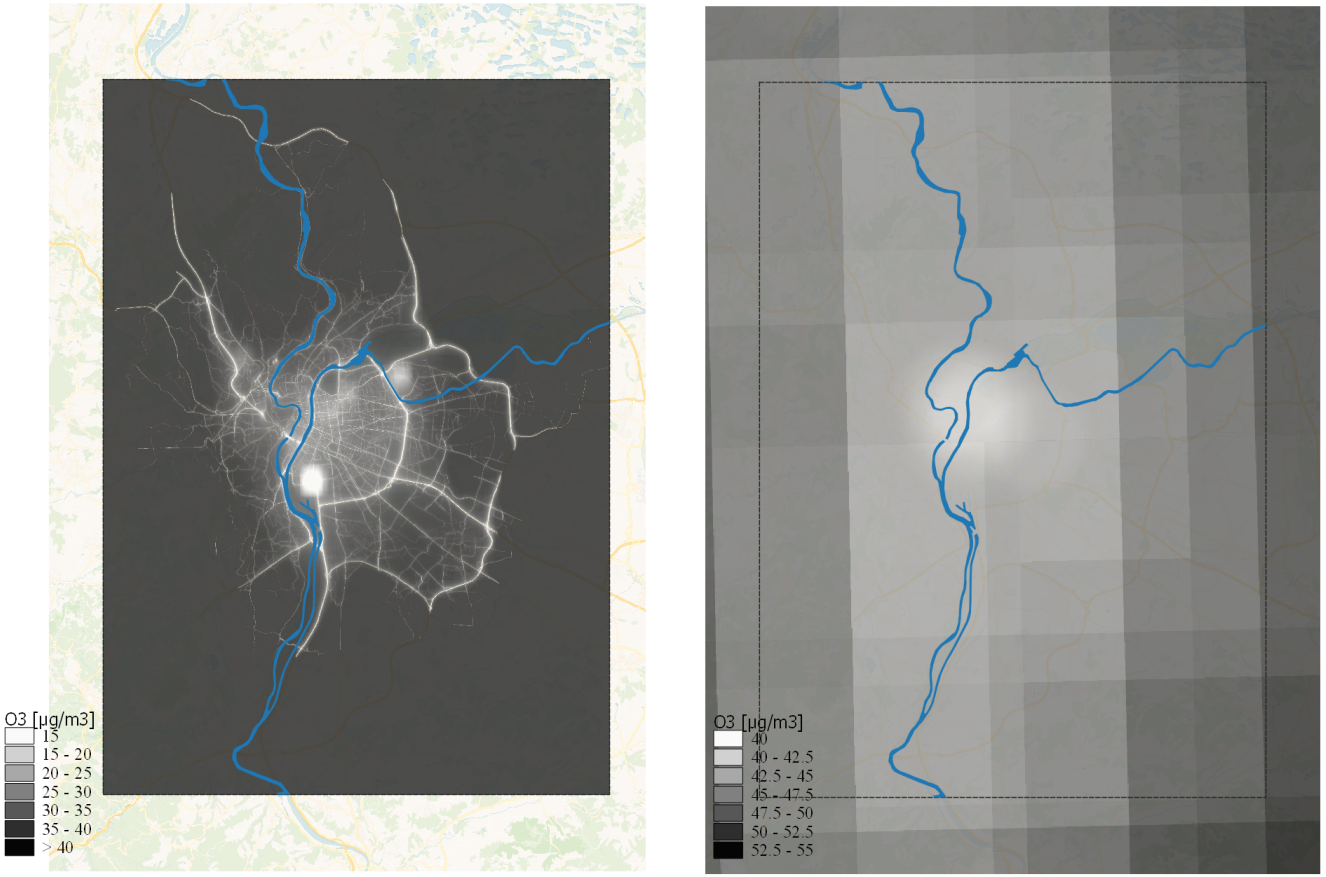

Figure C.5: O3 - 2000: SIRANE (left) and LUR (right)

Appendix D: CLC appendix

Level 1	Level 2	Level 3
1 Artificial surfaces	11 Urban fabric	111 Continuous urban fabric
		112 Discontinuous urban fabric
	12 Industrial, commercial and transport units	121 Industrial or commercial units
		122 Road and rail networks and associated land
		123 Port areas
		124 Airports
	13 Mine, dump and construction sites	131 Mineral extraction sites
		132 Dump sites
		133 Construction sites
	14 Artificial, non-agricultural vegetated areas	141 Green urban areas
		142 Sport and leisure facilities
2 Agricultural areas	21 Arable land	211 Non-irrigated arable land
		212 Permanently irrigated land
		213 Rice fields
	22 Permanent crops	221 Vineyards
		222 Fruit trees and berry plantations
		223 Olive groves
	23 Pastures	231 Pastures
	24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
		242 Complex cultivation patterns
		243 Land principally occupied by agriculture, with significant areas of natural vegetation
		244 Agro-forestry areas
3 Forest and semi natural areas	31 Forests	311 Broad-leaved forest
		312 Coniferous forest
		313 Mixed forest
	32 Scrub and/or herbaceous vegetation associations	321 Natural grasslands
		322 Moors and heathland
		323 Sclerophyllous vegetation
		324 Transitional woodland-shrub
	33 Open spaces with little or no vegetation	331 Beaches, dunes, sands
		332 Bare rocks
		333 Sparsely vegetated areas
		334 Burnt areas
		335 Glaciers and perpetual snow
4 Wetlands	41 Inland wetlands	411 Inland marshes
		412 Peat bogs
	42 Maritime wetlands	421 Salt marshes
		422 Salines
		423 Intertidal flats
5 Water bodies	51 Inland waters	511 Water courses
		512 Water bodies
	52 Marine waters	521 Coastal lagoons
		522 Estuaries
		523 Sea and ocean

Figure D.1: CLC nomenclature table

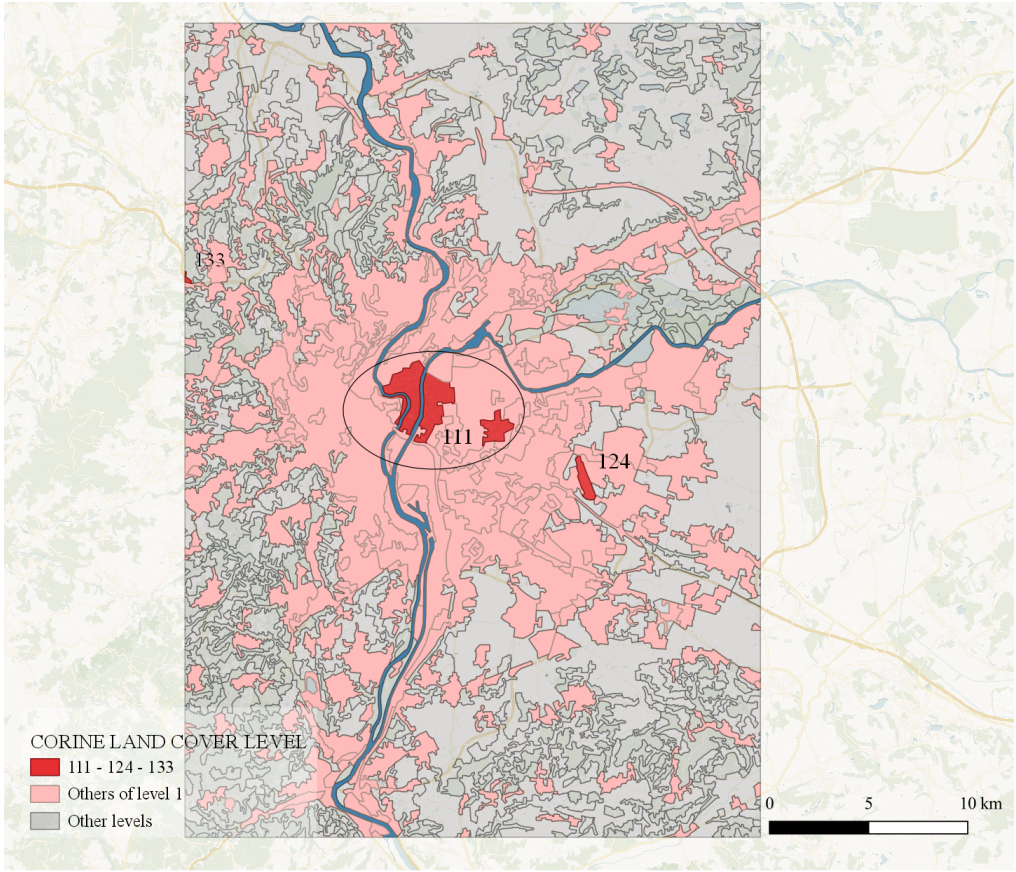

Figure D.2: CLC geometries: red ones are those with higher levels of misclassification for NO2 (111-124-133)

Appendix E: Example of odds ratio calculus methodology

The following figures represent a graphical explanation of the procedure explained in section 2.9 for the odds ratio calculation comparison. It is important to say that this appendix only contains a simplified example to understand the conceptual structure of the procedure, and consequently the figures and populations showed represent an extreme simplification of those involved in the study.

1- Definition of the exposure quartiles

The study domain (figure E.1, left) is divided in 4 surfaces (figure E.1, right), individuated by the quartiles of the distribution of population density-averaged exposure values computed by SIRANE for each cell of the computational domain.

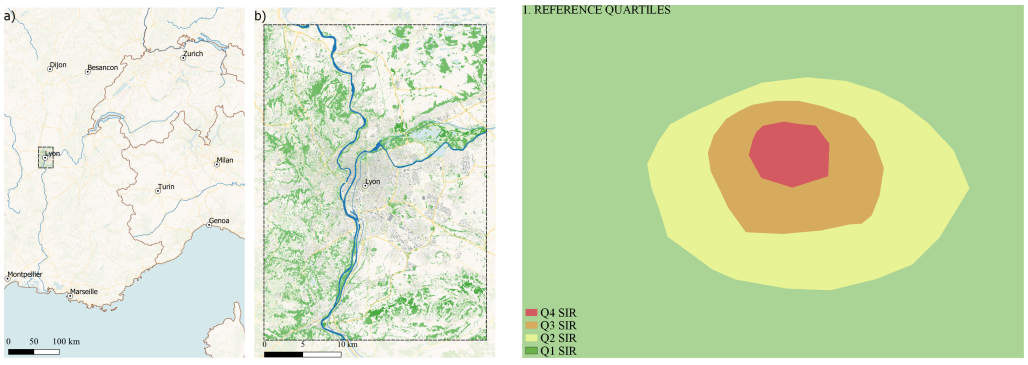

Figure E.1: Study domain and reference surfaces

2- Population and case-controls distribution

A population is created within the domain (figure E.2 on the left) so that the cases (white points) and controls (black points) distribution within the four surfaces is fixed as equal to a reference distribution, that in the main work is given by table 2.3 (in figure E.2 the risk is deliberately accentuated).

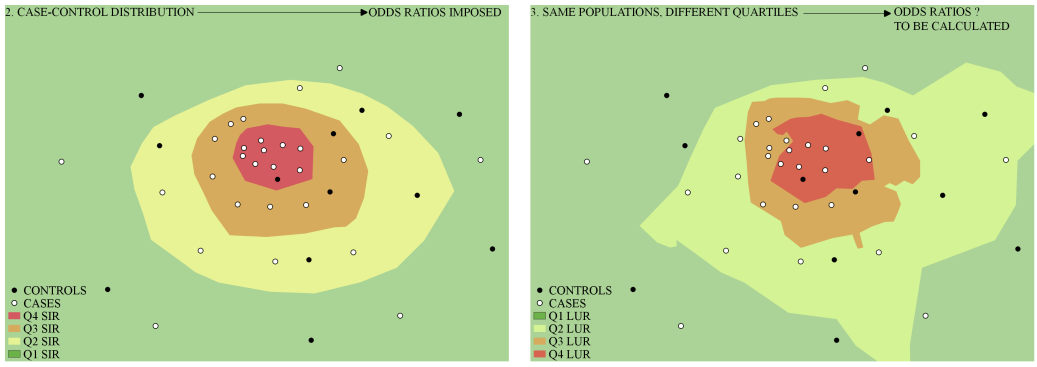

Figure E.2: Population distribution within SIRANE groups (reference, on the left) and LUR's (on the right)

3- Replacement with LUR surfaces and odds ratios calculation

In figure E.2 on the right, the distribution of the subjects within the quartiles is unavoidably changed because of the replacing of SIRANE values (that defined 4 surfaces) with LUR values, whose quartiles are different and so identifies four different surfaces. Odds ratios calculated are consequently different from the reference SIRANE scenario's ones.

4- Iteration

The procedure is repeated 500 times (figure E.3 is an example of the second iteration step), each time with a population whose cases and controls are equally distributed within SIRANE's surfaces (same amount of subjects in each surface, same division between cases and control in each surface, equal to the reference) and undergo a distribution modification when

passing at LUR's surfaces. New odds ratios are calculated each time: at the end, the averaged odds ratios resulting by the 500 simulations (with their averaged confidence interval extremes) are compared with the reference. Furthermore, since the reference distribution define significant odds ratios (all the 95% CI >1) for the third and fourth groups, the percentage of simulations that keep individuating the epidemiological significance for LUR's groups is calculated and discussed.

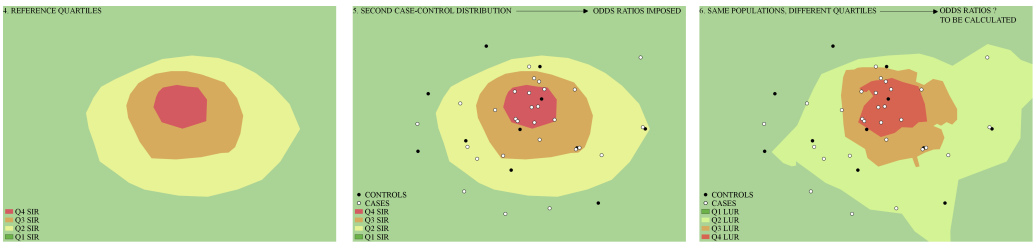

Figure E.3: Example of second iteration step